

to a wave of permanent form. A model equation describing the unidirectional propagation of long waves in water of relatively shallow depth with a localized solution representing a single hump as discovered by Russell was obtained by Korteweg and de Vries [42]. This equation has become very famous and is now known as the Korteweg–de Vries equation or KdV equation. These results were not considered very important when they were obtained. A remarkable numerical discovery was made by E. Fermi, J. Pasta and S. Ulam [18] as they studied the flow of incoherent energy in a solid modeled by a one-dimensional lattice of equal masses connected by nonlinear springs. The initial injected energy was not shared among all the degrees of freedom of the lattice, but returned almost entirely to the original excited mode. The explanation was given only ten years later by Zabusky and Kruskal [82] from numerical solutions of the KdV equation used to model in the continuum approximation a nonlinear atomic lattice with periodic boundary conditions. This led to the concept of solitons and ultimately to the development of integrable systems. The term soliton was chosen as it is a localized wave which propagates preserving its shape and velocity as a particle would propagate. A soliton is then a large-amplitude pulse of permanent form whose shape and speed are not altered by collision with other solitary waves, and it is the exact solution of a nonlinear equation. The mathematical features of solitons are often well developed in soliton literature since they are at the origin of very nice theoretical developments such as the powerful inverse scattering transform which solves a complex nonlinear equation through a series of linear steps. Nevertheless, the physics of solitons is very rich and it is a very actual research topic in numerous fields, for example in hydrodynamics, in optics, in electricity, in solid physics, in chemistry and in biology. Nonlinearity and dispersion are very common in macroscopic physics as well as in microscopic physics. Nonlinearity tends to localize the signals while dispersion spreads them. These two opposite effects sometimes compensate and the regime of solitons takes place. Real physical systems are only approximately described by the equations of the theory of solitons, but a remarkable feature of solitons is their very high stability relative to perturbations. The soliton concept is now firmly established. They are widely accepted as a structural basis for viewing and understanding the dynamic behavior of complex nonlinear systems.

## Introduction

The first experimental observation of a solitary wave was made in August 1834 by a Scottish engineer named John Scott Russell (1808–1882). He was mounted on a horse

along the Union Canal linking Edinburgh with Glasgow when he saw a rounded smooth well-defined heap detach itself from the prow of a boat brought to rest and continue its course without change of shape or diminution of speed. Scott–Russell reported his observation in the British Association Report [71] as follows:

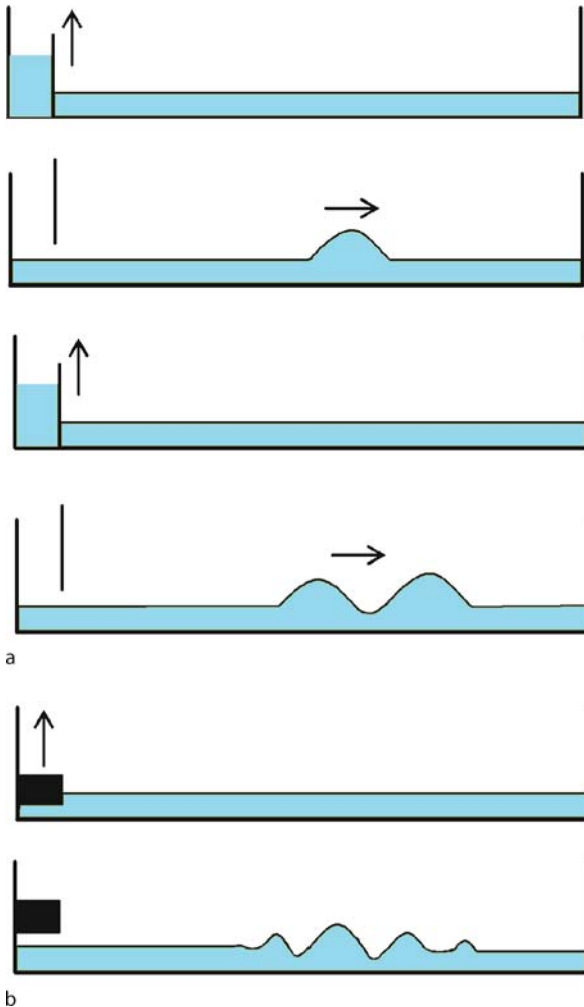
“I was observing the motion of a boat which was rapidly drawn along a narrow channel by a pair of horses, when the boat suddenly stopped—not so the mass of water in the channel which it had put in motion; it accumulated round the prow of the vessel in a state of violent agitation, then suddenly leaving it behind, rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed. I followed it on horseback, and overtook it still rolling on at a rate of some eight or nine miles an hour, preserving its original figure some thirty feet long and a foot to a foot and a half in height. Its height gradually diminished, and after a chase of one or two miles I lost it in the windings of the channel. Such, in the month of August 1834, was my first chance interview with that singular and beautiful phenomenon which I have called the Wave of Translation”.

Russell wrote that this August day in 1834 was the happiest in his life. In July 1995, an international gathering of scientists witnessed a re-creation of the solitary wave observed by Russell on an aqueduct of the Union Canal. This aqueduct was named after this occasion *Scott Russell aqueduct*; the scientists were attending a conference on nonlinear waves in physics and biology at Heriot-Watt University (Edinburgh), near the canal.

Throughout his life Russell remained convinced that “his” wave that he initially called the “Wave of Translation” and later the “Great Solitary Wave” was of fundamental importance. He therefore carried out extensive experiments in wave-tanks and established the velocity formula for solitary waves:

$$V_S = \sqrt{g(d + A_S)} \quad (1)$$

where  $g$  is the acceleration due to gravity,  $d$  the undisturbed water depth and  $A_S$  the amplitude of the solitary wave. The “quasi-cycloidal” form of the wave was found to be independent of the way it was generated. Russell [69] described the induced movement of fluid particles: “By the transit of the wave the particles of the fluid are raised from their places, transferred forwards in the direction of the motion of the wave, and permanently deposited at rest in a new place at a considerable distance from their original position”. He also deduced from his numerous experiments that two solitary waves can cross



**Solitons: Historical and Physical Introduction, Figure 1**  
Schematic view of Scott Russell's experiments

each other “without change of any kind”. Figure 1 shows a schematic view of his experiments in tanks, adapted from Remoissenet [65]. An initial elevation of water may induce one or two solitary waves depending on the relation between its height and length (Fig. 1a); the case where an initial depression does not evolve into solitary waves but leads to an oscillatory wave train of gradually increasing length and decreasing amplitude is shown in Fig. 1b.

In Sect. “[Historical Discovery of Solitons](#)”, the historical discovery of solitons is presented. The physical concept of solitons and the associated applications are described in Sect. “[Physical Properties of Solitons and Associated Applications](#)”. In Sect. “[Mathematical Methods Suitable for the Study of Solitons](#)”, the mathematical methods suitable for the study of solitons are considered. Finally, Sect. “[Future Directions](#)” is devoted to future directions.

### Historical Discovery of Solitons

After Russell had carried out his observation of the solitary wave in August 1834, he was put in charge by the Union Canal Company of determining the efficiency of canals for the transport of steam-driven barges. He worked on the resistance of boats towed through water of finite depth. A resistance proportional to the square of the velocity  $V_b$  of the boat was predicted by the theory at that time. Convinced of the imperfection of this theory, Russell performed numerous experiments in the summers of 1834 and 1835 [12]. Several boats were towed in canals of various depths at velocities ranging from 3 to 15 miles per hour. The resistance was measured by a dynamometer. Resistance curves such as the one depicted in Fig. 2 were obtained, with a resistance local maximum  $R_m$  for a boat velocity  $V_{cr}$  and a growing deviation from the theoretical curve for increasing values of boat velocity. The critical velocity  $V_{cr}$  was found to depend on the water depth. Russell noticed that this critical velocity was the same as the velocity of the solitary wave for the given water depth  $d$ , and gave a qualitative explanation for the diminution of flow resistance for increasing values of the boat velocity just above the critical velocity  $V_{cr}$ . For this description, he considered the variation of the shape of the water surface around the moving boat with the velocity. For velocities smaller than the critical value, the water level is raised around the prow where a wave of displacement is generated, leading to an inclination of the boat. The resulting effect is an augmentation of the section of immersion and of the corresponding flow resistance. When the velocity of the boat reaches the critical value, the wave of displacement has the velocity of a solitary wave and the vessel travels on this solitary wave. For velocities greater than the critical value, the boat stays on the wave summit leading to a smaller flow resistance. Russell [70] tried a quantitative explanation of this process, but this paper contained several errors. However, Russell was a very fine observer; he was far ahead of his time in considering the fundamental importance of solitary waves. He mentioned [70] that “a large or high wave had a greater velocity than a small one. When a small wave preceded a large one, the latter invariably overtook the other, and when the large wave was before the less, their mutual distance invariably became greater”. He showed that the shape of solitary waves depends on their height, the width of the wave decreasing for increasing values of height. It is also fascinating that Russell had observed the collision properties of solitary waves [8]. Nevertheless, the observations of Russell induced numerous critical comments. Airy [2] wrote in his treatise “Tides and Waves”: “We are not disposed to recognize this wave (discovered by Scott

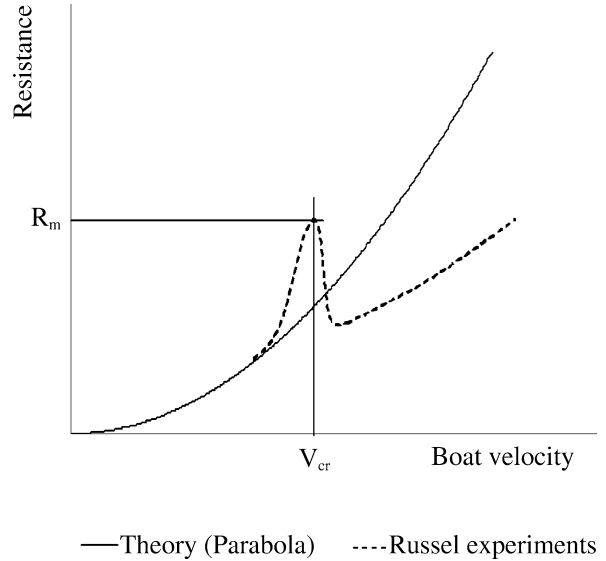
Russell) as deserving the epithets “great” or “primary”, and we conceive that ever since it was known that the theory of shallow waves of great length was contained in the equation  $\partial^2 X/\partial t^2 = g d \partial^2 X/\partial x^2$  the theory of the solitary wave has been perfectly well known”. In the equation quoted by Airy,  $t$  is the time and  $x$  is the horizontal coordinate of a fluid particle. The wave velocity in shallow water at the first order of approximation, when the ratio  $H/d$  between the wave height  $H$  and the water depth may be neglected is  $V_0 = \sqrt{gd}$ , as found by Lagrange in 1786; this velocity differs from the formulation proposed by Russell for solitary waves (Eq. (1)). Airy mentioned further: “Some experiments were made by Mr. Russell on what he calls a negative wave. But (we know not why) he appears not to have been satisfied with these experiments and had omitted them in his abstract. All of the theorems of our IVth section, without exception, apply to these as well as to positive waves, the sign of the coefficient only being changed”. Also, according to Airy [2] and Lamb [46], long waves could not propagate in a canal of rectangular section without changing their shape. The controversy arose because the dispersion is neglected by the nonlinear shallow water theory. Stokes [73] considered Russell results and tried to obtain them analytically; however, his linear or weakly nonlinear approach did not permit him to retrieve the results of Russell, which made him doubtful about them. In order to distinguish his “great wave” from other waves, Russell [71] introduced four orders of waves:

1. Waves of translation. These waves involve mass transfer. The solitary waves observed by Russell are included in this order.
2. Oscillatory waves. These are the waves which can be the most often observed; they do not involve mass transfer.
3. Capillary waves. The surface tension effects are important for those waves.
4. Corpuscular waves. Those waves are rapid successions of solitary waves.

Russell was most concerned with the first order, but he also carried out experiments with waves of the second and third order. The fourth order (corpuscular waves) suggested by Russell had been seriously criticized, and the manuscripts he submitted on this order were never published due to ignorance of mechanics principles.

The great difficulties encountered by Russell to prove the importance of his findings strongly disappointed him. Tired of discussions, he stopped his work on solitary waves and started to build large steam ships with great success. A detailed biography of Russell may be found in [16].

A French mathematician, Joseph Boussinesq, knew of the existence of Russell’s observations, and also of detailed



Solitons: Historical and Physical Introduction, Figure 2

Resistance as a function of towing velocity according to Russell

experiments performed by Bazin [3] in the long branch of the canal de Bourgogne close to Dijon (France), which confirmed Russell’s observations. Boussinesq tried to obtain a solution of Euler’s equations compatible with Russell’s and Bazin’s results. He developed the horizontal and vertical velocity components of fluid velocity in a rectangular channel in power of the distance from the bed. Including nonlinear terms which had been neglected by Lagrange, Boussinesq [5] obtained a solution with the properties of the solitary wave observed by Russell. In particular, the shape of the wave was found to be given by:

$$\eta = A_S \operatorname{sech}^2 \left[ \sqrt{\frac{3A_S}{4d^3}} (x - V_S t) \right] \quad (2)$$

where  $\eta$  is the free surface displacement and  $\operatorname{sech} x = 1/\cosh x$ ; the velocity of the wave  $V_S$  is given by Eq. (1). The article of Boussinesq [5] had been almost ignored in England, and five years after its publication Lord Rayleigh [64] independently obtained the solitary wave profile. Following Rayleigh’s work and including the effects of surface tension, the Amsterdam professor of mathematics Diederik Johannes Korteweg and his doctoral student Gustav de Vries derived a model equation [42] which describes the unidirectional propagation of long waves in water of relatively shallow depth. This famous equation is now known as the Korteweg–de Vries equation or KdV equation, and it has the following form:

$$\frac{\partial \eta}{\partial t} + V_0 \frac{\partial \eta}{\partial x} + \frac{3}{2} \frac{V_0}{d} \eta \frac{\partial \eta}{\partial x} + \frac{1}{6} V_0 d^2 \frac{\partial^3 \eta}{\partial x^3} = 0. \quad (3)$$

Korteweg and de Vries showed that this nonlinear wave equation has a localized solution which represents a single hump of positive elevation corresponding to the solitary wave observed by Scott Russell. The KdV equation had been in fact formulated in 1877 by Boussinesq in an impressive paper [6], but this author did not use directly this formulation. Russell's experiments had found their theory and one could think that many scientists would rapidly extend the results of this theory. However, the KdV equation had a quiet life for many decades.

In the early 1950s, the physicists Enrico Fermi, John Pasta and Stan Ulam could use one of the first computers, the "Maniac" to work on systems without closed analytic solutions. In particular, they studied the way a crystal evolves towards the thermal equilibrium through the numerical simulation of the dynamics of a one-dimensional lattice consisting of  $N$  equal masses (atoms) connected by nonlinear springs. The problem is described by the Hamiltonian

$$H = \sum_{i=0}^{N-1} \frac{1}{2} p_i^2 + \sum_{i=0}^{N-1} \frac{1}{2} K(x_{i+1} - x_i)^2 + \frac{K\alpha}{3} \sum_{i=0}^{N-1} (x_{i+1} - x_i)^3 \quad (4)$$

where  $x_i$  is the atom  $i$  displacement,  $p_i$  the corresponding momentum,  $K$  the constant of the quadratic potential, and  $\alpha (\ll 1)$  a coefficient of nonlinear interaction. The boundary condition  $x_0 = x_N = 0$  is assumed. Considering a normal mode decomposition

$$A_k = \sqrt{\frac{2}{N}} \sum_{i=0}^{N-1} x_i \sin(ik\pi/N), \quad (5)$$

the Hamiltonian (Eq. (4)) may be written in the following way

$$H = \frac{1}{2} \sum_k \left( \dot{A}_k^2 + \omega_k^2 A_k^2 \right) + \alpha \sum_{k,l,m} c_{klm} A_k A_l A_m \quad (6)$$

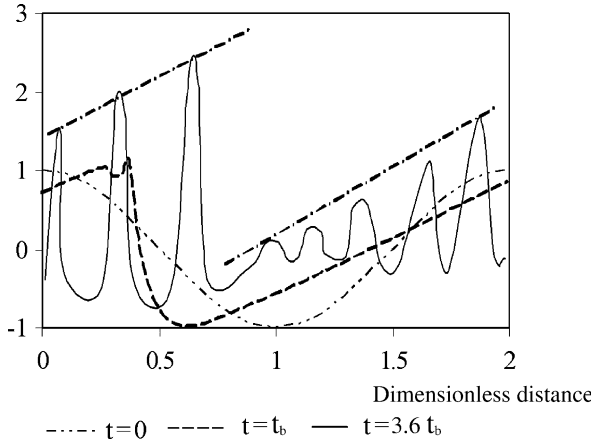
where the pulsations  $\omega_k$  are given by  $\omega_k^2 = 4K \sin^2(k\pi/2N)$ , and  $c_{klm}$  are constants. When the masses have very small displacements from the equilibrium positions, the dynamics of the lattice may be described by a superposition of normal modes. It was supposed that if only one mode (the fundamental mode  $k = 1$  in the circumstances) is excited in a crystal lattice, nonlinear interactions would lead to energy equipartition or thermal equilibrium. This was suggested by the results at the beginning of the calculations since the modes 2, 3, ... were successively excited. However, after 157 fundamental periods, almost all

the energy returned to the fundamental mode. Furthermore, the energy returned almost periodically to the original excited mode leading to a near recurrence process, now known as the FPU paradox. Fermi et al. [18] did not get the expected result and the modeled crystal lattice did not approach energy equipartition. This remarkable discovery is at the origin of the comprehension of the concept of the soliton and of the developments of "numerical experiments". Fermi died just after this work; this had a rather inhibitory effect on the significance of this result and no paper was published after the preprint "Studies of nonlinear problems" [18]. The results were communicated to the scientific community by Ulam through several conferences. Complementary studies [20,39] confirmed that the introduction of nonlinearity in a system does not guarantee the equipartition of energy. In 1960, the KdV equation was rederived by Gardner and Morikawa who worked on collisionless hydromagnetic waves. These authors attracted the attention of Zabusky and Kruskal on the KdV equation, and the FPU paradox was solved a few years later in terms of solitons [82]. Zabusky and Kruskal considered the following motion equations corresponding to the Hamiltonian (Eq. (4))

$$\ddot{x}_i = K(x_{i+1} + x_{i-1} - 2x_i) + K\alpha[(x_{i+1} - x_i)^2 - (x_i - x_{i-1})^2]. \quad (7)$$

Taking into account that the recurrence occurs before the excitation of high-order modes, these authors carried out an asymptotic analysis in the continuum approximation and obtained an equation including dispersive and nonlinear terms which correspond to the KdV equation. This equation, obtained in a totally different physical context as the one to explain the observations of Scott Russell, is at the origin of the explanation of the FPU paradox. As depicted in Fig. 3, the numerical experiments of Zabusky and Kruskal showed that the initial sinusoidal condition evolves into steep fronts and then into a finite number of short pulses which are solitons. These solitons travel around the lattice with velocities depending on their amplitude; they collide preserving their individual shapes and velocities with only a small change in their phases. As time increases, there is an instant at which the solitons collide at the same point, and the initial state comes close to recurrence. The recurrence period calculated by Zabusky and Kruskal closely approximates the actual FPU recurrence period, showing that the KdV equation is a suitable approximation of the FPU system. The system had to be considered as totally nonlinear to be solved, and the linear normal modes of the system could not be considered. The pulse-like waves were called solitons by Zabusky and





**Solitons: Historical and Physical Introduction, Figure 3**

Evolution of an initially periodic profile from Zabusky and Kruskal numerical analysis of the KdV equation. The breaking time for the wave profile is  $t_b$

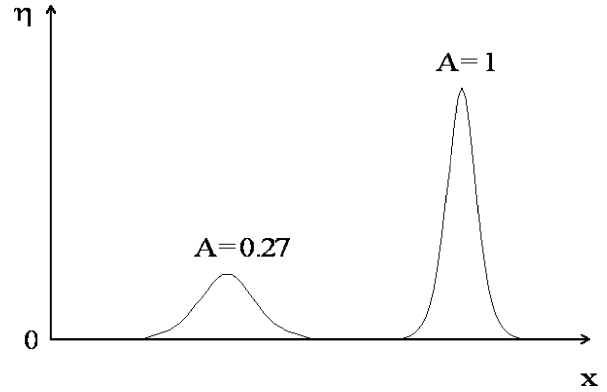
Kruskal to indicate the remarkable quasi-particle properties of these very stable solitary waves. They were at first called solitons to introduce them into the family of particle names such as electron and neutron, but the name soliton was a trade mark, and was therefore not suitable. The concept of solitons has spread over numerous branches of physics, such as optical physics, biological physics, astronomy, particle physics, condensed matter, fluid dynamics, ferromagnetism. The numerical experiments of Zabusky and Kruskal led to the development of integrable systems. Taniuti and Wei [76] and Su and Gardner [74] have shown that a large class of nonlinear evolution equations may be reduced to consideration of the KdV equation which may be regarded as a very important canonical form. A famous method, the so-called Inverse Scattering Transform (IST) was proposed by Gardner et al. [24] to integrate the KdV equation. This procedure may be considered as the nonlinear analogue of the Fourier transform method suitable to linear dispersive systems.

### Physical Properties of Solitons and Associated Applications

#### Properties of Solitons

Let us consider the KdV equation (Eq. (3)). Introducing the variable  $X = x - V_0 t$ , this equation may be written in the following way

$$\frac{1}{V_0} \frac{\partial \eta}{\partial t} + \frac{3}{2d} \eta \frac{\partial \eta}{\partial X} + \frac{d^2}{6} \frac{\partial^3 \eta}{\partial X^3} = 0. \quad (8)$$



**Solitons: Historical and Physical Introduction, Figure 4**

Comparison between two solutions of the KdV equation with different amplitude  $A$

Using the dimensionless variables  $\phi = \eta/d$ ,  $\xi = X/X_0$  and  $\tau = t/t_0$  where  $X_0$  and  $t_0$  are respectively a typical length and time, the KdV equation may be formulated in its standard form:

$$\frac{\partial \phi}{\partial \tau} + 6\phi \frac{\partial \phi}{\partial \xi} + \frac{\partial^3 \phi}{\partial \xi^3} = 0. \quad (9)$$

This equation has many solutions, as all nonlinear equations. Among these solutions, there are the spatially localized solutions:

$$\phi = A \operatorname{sech}^2 \left[ \sqrt{\frac{A}{2}} (\xi - 2A\tau) \right], \quad (10)$$

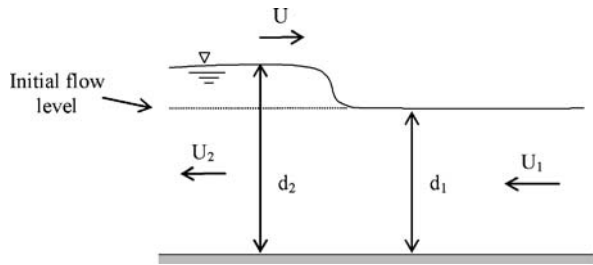
where  $A$  is a positive constant. These solutions quantitatively correspond to the observations of Scott Russell. In particular, the width of the soliton decreases for increasing values of its amplitude, as depicted in Fig. 4. In dimensional variables and in a fixed referential, the corresponding solution is given by Eq. (2).

The KdV equation appears widely in physics, each time that waves propagate in a weakly nonlinear and weakly dispersive medium. The nonlinear term  $\phi(\partial\phi/\partial\xi)$  in this equation causes the steepness of the wave form. In a dispersive medium, the Fourier components of a pulse propagate at different velocities, inducing a spreading of the pulse. The dispersive term  $\partial^3\phi/\partial\xi^3$  in the KdV equation makes the wave form spread. The soliton solution of the KdV equation results from the balance of nonlinearity and dispersion. This balance is generally very stable, which explains the numerous applications of the theory of solitons, even if the real physical situations (where friction and defects take place) are only approximately described by the soliton equations, and in particular by the KdV

equation. It would be strictly speaking more appropriate to use the terminology quasi-soliton in physics; however, bearing in mind that the word soliton has a quite ideal connotation in the context of systems with exact solutions, we will use the word soliton. The conservation of the shape and of the velocity of the soliton after a collision is a manifestation of stability. The high stability of solitons relative to perturbations, combined with the fact that they can spontaneously emerge in a physical system in which energy is supplied, lead to numerous applications of solitons in macroscopic and microscopic physics. Nevertheless, a soliton may sometimes vanish. For example, this is the case when such a wave propagates in a media where the water depth decreases during the wave propagation. The dispersive effect progressively decreases when the nonlinear effect increases, leading to the wave breaking as waves on a beach. The vanishing of a soliton may also result from energy dissipation. Marin et al. [52] have shown that sand ripples may form when solitons propagate in shallow water over a sandy bed, and that a strong interaction between the free surface and the bed occurs. This interaction leads to a decrease of the soliton amplitude, which may result under certain circumstances in the disappearance of the soliton. Let us now consider some of the most typical soliton applications.

### Solitons in Fluid Mechanics

In the hydrodynamic field, a very nice case where solitons take place is the tidal bore, also called hydraulic jump in translation, and known as the mascaret in France. A tidal bore is a positive wave of translation which can occur in a river with a gently sloping bottom and a broad funnel-shaped estuary, where a difference of more than six meters between high and low water takes place [50]. The tidal bore appears as a single wave or a series of waves which propagate upstream. At a given instant, it is localized at the upstream extremity of the flood tide propagation in the inter-tidal zone. The flow direction is towards the sea before the tidal bore arrival, and may be in the opposite direction after its occurrence. Such bores are distributed widely throughout the world. The most powerful can be seen in the Amazon river in South America, where it is called pororoca, which means gigantic noise. An approximately 5 m-high wave propagates with a velocity of about 30 m/s, and surfers take advantage of this exceptional wave to organize competitions. In England, bores may be contemplated on several rivers, the most famous occurring in the Severn river, near Bristol. In France, bores take place mainly in the Mont Saint-Michel bay, in the Gironde, Dordogne and Garonne rivers. The mascaret of the Seine river



**Solitons: Historical and Physical Introduction, Figure 5**  
Sketch of the propagation of a positive wave of translation

almost disappeared between 1960 and 1970 because of the construction of dikes and the dredging works which had been carried out in order for large boats to reach the city of Rouen. Increasing the water depth induces a decrease of the nonlinear effects which become insufficient to balance the dispersion. In Canada, several bores are observed in Fundy bay. In Asia, an important bore occurs in the Qiantang river, while the one in the river Pungue in Africa can propagate upstream over more than 80 km. Two types of bores may be observed: the breaking bore and the undulated bore, depending on the value of the Froude number  $Fr$  defined by  $Fr = (U_1 + U)/\sqrt{gd_1}$ , where  $U_1$  is the flow velocity at the depth  $d_1$  before the passage of the bore and  $U$  the bore velocity (Fig. 5). In Fig. 5,  $d_2$  depicts the water depth after the passage of the bore and  $U_2$  the flow velocity at the depth  $d_2$ . For a value of  $Fr$  lower than about 1.5, the bore is undulated, and for a value of  $Fr$  greater than 1.5, the breaking bore occurs. Most of the bores are undulated bores. The tidal bore is a very important turbulent phenomenon for the inter-tidal zone of a river, inducing a significant mixing of waters and complex motions of sediment on the river bed [14]. River bores have been known for many centuries but only qualitative observations were carried out. Quantitative measurements came much later, mainly because solitons are inherently nonlinear when only linear processes were considered.

Long waves such as tsunamis often behave like solitary waves [48]. In particular, the run-up and shoreward inundation are often simulated using solitary waves [49]. The tsunamis may be described by the KdV equation. The theory of very long shallow water waves is valid since the tsunamis have a wavelength which is generally greater than 100 km and they propagate in oceans whose mean depth is about 4000 m. Their propagation velocity is greater than  $V_0 = \sqrt{gd}$ , that is greater than approximately 700 km/h. When they approach a coast, their velocity decreases and their height increases to satisfy the equation of energy conservation, and they finally break inducing often

very significant damages on the shore. Most tsunamis have a seismic origin, such as the dramatic one which occurred in the Indian Ocean on 26 December 2004 and which led to the death of more than 220 000 people. These solitons may also result from a quick arrival in the sea of volcanic products, through a similar process to that in Scott Russell's experiments depicted in Fig. 1. This was the case for the tsunami generated in 1883 by the eruption of Krakatau in Indonesia, when approximately 36 000 people died. The inundation phase for a tsunami highly depends on the bottom bathymetry. The breaking can be progressive and structural damages are mainly caused by inundation [13]. The breaking can also be explosive and induce the formation of a plunging jet. When the breaking takes place in very shallow water, the tsunami amplitude becomes so high that an undulation appears on the long wave, which develops into a progressive bore [10]. This turbulent front is of the same type as the wave which occurs when a dam breaks, and the water can rise very rapidly, typically from 0 to 3 meters in 1.5 minutes. The event of December 2004 has shown that the available models are not able to predict accurately the wave run-up heights under severe conditions.

Helal and El-Eissa [37] presented the connection between shallow water waves and the KdV equation. More recently, Mei and Li [55] considered the propagation of long waves over a randomly rough seabed. These authors have shown analytically and numerically that, assuming a randomness height to mean depth ratio comparable to the one of mean depth to the characteristic wavelength, disorder causes diffusion, leading to spatial attenuation of the wave amplitude. Mei and Li [55] proposed a modified KdV equation which includes terms representing the effects of disorder on amplitude attenuation. After a propagation over a region of finite length, the transmitted wave is a pulse which disintegrates into several small solitons of vanishing energy.

Hydrodynamic solitons may also be observed in deep water. Stokes showed in 1847 through the use of a small amplitude expansion of a sinusoidal wave, that periodic waves of finite amplitude are possible in deep water. The Stokes waves are unstable to infinitesimal modulation perturbations [4]. Zakharov and Shabat [83] were the first to show that an initial wave packet may evolve into a number of envelope solitons and a dispersive tail, the envelope solitons consisting of a carrier wave modulated by an envelope signal. This was experimentally verified by Yuen and Lake [81], by carrying out detailed experiments in deep water in a wave tank. The stability properties of envelope solitons were also considered. The freak waves, also called rogue waves, extreme or giant waves, may also have

soliton-like shape. These waves appear surprisingly in the sea ("wave from nowhere") in deep or shallow water, and can cause severe damages to ships and fixed ocean structures. They are characterized by their brief occurrence in space and time, resulting from a local focusing of wave energy [19]. The main features of their physical processes may be obtained performing numerical simulations in the framework of classical nonlinear evolution equations, such as the KdV equation [41].

Oceanographers have also observed internal solitary waves in many regions around the world's oceans [61]. Most of the observed solitons propagate at the interface between the thermocline and the deep sea [56]. They are mainly excited by tidal flows over bottom topography [25], and they contribute to significant vertical mixing. Brandt et al. [7] developed a rotationless Boussinesq-like model for the generation and propagation of internal waves; the results of the simulations are in good agreement with large solitary waves recorded with airborne Radar images. Michallet and Barthélemy [56] carried out experiments in a 3-m-long flume to study interfacial long-waves in a two-immiscible-fluid system involving water and petrol. The comparison between the experiments and nonlinear theories shows that the KdV solitary waves match the experiments for small amplitude waves for all layer thickness ratios. When the crest is close to the critical level, that is approximately located at mid-depth, the large amplitude waves are well predicted by a modified KdV equation including both quadratic and cubic nonlinear terms (KdV-mKdV equation). However, only very few laboratory or field measurements of internal solitary waves are available, due to the prohibitive cost of obtaining precise measurements in the field and to the difficulty of generating and measuring these waves in the laboratory.

Blood pressure pulses may be considered as solitons [63]. There is a balance between the nonlinearity due to the blood flow and the dispersion connected to the elasticity of the artery. Previous studies [32,33,80] have shown that the dynamics of weakly nonlinear pressure waves in a thin nonlinear elastic tube filled with an incompressible fluid may be governed by the KdV equation. A Boussinesq-like equation was obtained by Paquerot and Remoissenet [62] to describe the blood pressure propagation in large arteries, in the limit of an ideal fluid and for slowly varying arterial parameters.

Several methods have been tested to generate hydrodynamic solitons in a flume; the method used by Scott Russell has been described in Sect. "Introduction" (Fig. 1). Hammack and Segur [29] showed experimentally and theoretically that from any net positive volume of water above the still water level, at least one solitary wave will emerge fol-

lowed by a train of dispersive waves. Maxworthy [54] and Weidman and Maxworthy [79] pushed the liquid in the horizontal direction by a single displacement of a piston. The generation of solitary waves using a piston-type wave maker has been studied in detail by Goring [26]. Guizien and Barthélemy [28] have proposed an experimental procedure to generate solitary waves in a flume using a piston type wave maker from Rayleigh's (1876) solitary wave solution. The advantage of this method is that solitary waves are rapidly established with very little loss of amplitude in the initial stage of the propagation of the solitary waves. The generation of solitons using a wave flume in resonant mode, without an absorbing beach, has been considered by Marin et al. [52]. Surface waves are produced in shallow water by an oscillating paddle at one end of the flume; a near-perfect reflection takes place at the other end. The frequency of the oscillating paddle was chosen close to the resonant frequency of the mode whose wavelength is equal to the flume length. For small values of the amplitude of displacement of the oscillating paddle, only standing harmonic waves are generated in the flume. For values of this amplitude greater than a critical value, pulses propagating from one end of the channel to the other end are excited on the background of the standing harmonic wave. From one to four pulses propagating in each direction of the flume could be generated over the time period of the flow, depending on the frequency and the amplitude of horizontal displacement of the oscillating paddle. These pulses were identified as solitons (one pulse) and bound states of solitons (several pulses). The generation of solitons in the flume results from the excitation of high harmonics. The spatiotemporal properties of solitons generated in this way were studied in detail by Ezersky et al. [17]. Space-time diagrams have been constructed to highlight the spatiotemporal dynamics of nonlinear fields for two solitons colliding in the resonator. Period doublings and the multistability of the nonlinear waves, i. e. the occurrence of different regimes for the same values of the control parameters but under different initial conditions, have been shown. It is important to keep in mind that the solitary waves correspond to the limit of cnoidal waves of period tending to infinity. The velocity of cnoidal waves  $V_{cn}$  is known to depend on the so-called elliptic parameter  $m$  and is given by the formula [11]

$$V_{cn} = \left[ 1 - \frac{A_s}{2d} + \frac{A_s}{md} \left( 1 - \frac{3E(m)}{2K(m)} \right) \right] \sqrt{gd} \quad (11)$$

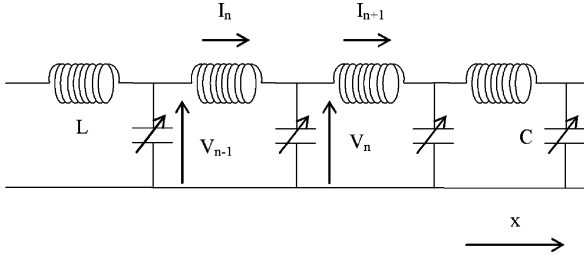
where  $K(m)$  and  $E(m)$  are complete elliptic functions of the first and second kind. The parameter  $m$ , which depends on the wave amplitude to width ratio, is responsible for the shape of the wave with  $m = 0$  corresponding to the

harmonic wave, and  $m = 1$  to the soliton. Strictly speaking, the waves generated in a flume used in resonant mode are not exactly solitons, but cnoidal waves with an elliptic parameter very close to 1; the value of this parameter was found to be 0.9996 for the tests carried out by Ezersky et al. [17]. The shapes of pulses are almost indistinguishable for  $m = 1$  and  $m = 0.9996$ , but the pulse repetition rates differ appreciably, the period tending to infinity as  $m \rightarrow 1$ .

Atmospheric solitons also exist, such as the Morning Glory Cloud of the Gulf of Carpentaria in northern Australia, which can be seen as a spectacular propagating roll cloud. Atmospheric solitons are horizontally propagating nonlinear internal gravity waves that can travel over large distances with minimal change in form [68]. The solitary waves that have been observed in the atmosphere can be divided into two classes: those that propagate in a fairly shallow stratified layer near the ground and those that occupy the entire troposphere. The generation mechanisms appear to be quite different for these two classes of waves. The waves that occupy the lower part of the troposphere mainly involve a gravity current such as a thunderstorm outflow, a katabatic wind, a sea breeze front, or a downslope windstorm, interacting with a low-level stable layer. The motion of the gravity current produces perturbations that are trapped in the low-level stable layer and eventually evolve over time into a series of solitary waves. For the tropospheric scale waves, they are very probably generated by synoptic scale features such as large-scale convective systems and geostrophic adjustments. For the two types of atmospheric solitons, there must exist some feature in the atmosphere that serves to prevent the wave energy from propagating away in the vertical direction. These trapping mechanisms are either deep layers of the atmosphere with very low values of the buoyancy frequency, or critical layers. The KdV equation, eventually enhanced with higher-order nonlinearity, gives a reasonable agreement with the observations [68].

### Solitons in Nonlinear Transmission Lines

In another area of physics, transmission lines with nonlinear elements are found to propagate in a soliton mode. Nonlinear electrical transmission lines are simple experimental devices to observe and consider quantitatively the propagation and properties of solitons. The propagation of these waves is used for picosecond impulse generation and broadband millimeter-wave frequency multiplication [66]. Let us consider the elementary LC network depicted in Fig. 6, with linear inductors  $L$  and nonlinear capacitors  $C$ . The differential capacitance  $C(V_n)$  is supposed to depend nonlinearly on the voltage across the  $n$ th



**Solitons: Historical and Physical Introduction, Figure 6**

Elementary circuit of an electrical network with linear inductance's  $L$  and nonlinear capacitance's  $C$

capacitor  $V_n$ :

$$C(V_n) = \frac{dQ_n(V_n)}{dV_n} \quad (12)$$

where  $Q_n(V_n)$  is the charge stored in the  $n$ th capacitor. Following Kirchhoff's law, we have:

$$V_{n-1} - V_n = \frac{d\phi_n}{dt}, \quad I_n - I_{n+1} = \frac{dQ_n}{dt} \quad (13)$$

where the magnetic flux  $\phi_n$  is related to the current  $I_n$  by the relation:  $\phi_n = LI_n$ . Using Eqs. (12) and (13), it is easy to obtain:

$$\frac{d^2 Q_n}{dt^2} = \frac{1}{L}(V_{n+1} + V_{n-1} - 2V_n), \quad n = 1, 2, \dots \quad (14)$$

Assuming the following capacitance-voltage relation,

$$C(V_n) = C_0(1 - 2aV_n) \quad (15)$$

where  $a$  is a small nonlinear coefficient, we find:

$$LC_0 \frac{d^2 V_n}{dt^2} - LC_0 a \frac{d^2 V_n^2}{dt^2} = (V_{n+1} + V_{n-1} - 2V_n), \quad n = 1, 2, \dots \quad (16)$$

The system (Eq. 16) of nonlinear equations cannot be solved analytically. The continuum limit is used to get approximate solutions; setting the position  $x = n\delta$  where  $\delta$  is a hypothetical cellule length, we get

$$\frac{\partial^2 V}{\partial t^2} - \frac{\delta^2}{LC_0} \frac{\partial^2 V}{\partial x^2} = \frac{\delta^4}{12LC_0} \frac{\partial V}{\partial x^4} + a \frac{\partial^2 V^2}{\partial t^2}. \quad (17)$$

This weakly dispersive and nonlinear wave equation describes waves that can travel both to the left and to the

right; the dispersive nature is due to the discreteness of the electrical network. A localized wave solution of the Eq. (17) that does not change its shape as it propagates with constant velocity  $v$  may be found [65]:

$$V = \frac{3(v^2 - v_0^2)}{2av^2} \operatorname{sech}^2 \left( \frac{\sqrt{3(v^2 - v_0^2)}}{v_0} \left( n - \frac{v}{\delta} t \right) \right) \quad (18)$$

where  $v_0 = \delta/\sqrt{LC}$ . This solitary wave solution represents a pulse with amplitude

$$V_m = \frac{3}{2a} \frac{v^2 - v_0^2}{v^2} \quad (19)$$

which depends on the velocity  $v$ . The width  $L_w$  at half height of this pulse is given by

$$L_w = 1.76 \frac{v_0}{\sqrt{3(v^2 - v_0^2)}}. \quad (20)$$

The waveforms of this solitary wave and of the KdV soliton are similar, and the width  $L_w$  depends on the amplitude  $V_m$ :

$$V_m L_w^2 = C^t \quad (21)$$

where  $C^t$  is a constant.

### Solitons in Plasmas

Solitons may propagate in plasmas; this has received much attention because of a possible relevance to final state configurations in fusion devices. The combination of dispersion and nonlinearity in plasmas may be described by the KdV equation. A direct analogy between the equations for shallow water and those for plasmas has been mentioned by Gardner and Morikawa [23]. However, the KdV equation is not always adapted to the excitations which can be generated in plasmas by electromagnetic waves, since wave packets are often produced instead of pulses characterizing the solutions of the KdV equation. Another type of soliton equation is then more suitable, the nonlinear Schrödinger equation (NLS); this second class of soliton equation will be considered in Subsect. "Solitons in Optical Fibers".

### Solitons in a Chain of Pendulums

Hydrodynamic solitons, solitons in nonlinear transmission lines are nontopological solitons, because the system returns to its initial state after the passage of the wave. There is also another type of soliton, the kink solitons which are topological solitons, since the structure of



the system is sometimes modified after the passage of the wave. A mechanical system consisting of a chain of pendulums, each pendulum being elastically connected to its neighbors by springs, is a typical example for which topological solitons can merge. It is easy to show that this mechanical system may be described by the following equation:

$$\frac{\partial^2 \theta}{\partial t^2} - c_0^2 \frac{\partial^2 \theta}{\partial x^2} + \omega_0^2 \sin \theta = 0 \quad (22)$$

where  $\theta$  is the angle of rotation of the pendulums,  $x$  is the axis along which the pendulums are distributed,  $c_0^2 = l_1^2 \beta / I_i$ ,  $l_1$  being the distance between two pendulums,  $\beta$  the torque constant of a section of spring between two pendulums,  $I_i$  the moment of inertia of a single pendulum of mass  $m$  and length  $l_2$ , and  $\omega_0^2 = mg l_2 / I_i$ . The first term in Eq. (22) represents the inertial effects of the pendulums, the second term corresponds to the restoring torque due to the coupling between pendulums, and the third term represents the gravitational torque. The Eq. (22) is called the Sine–Gordon (SG) equation [47]; it can be totally integrated and admits exact solitons solutions. It has become a very famous equation containing dispersion and nonlinearity which is used to model various phenomena in physics. It constitutes the third class of nonlinear equation leading to solitons.

### Fluxons in a Josephson Tunnel Junction

Solitons arise also in a Josephson tunnel junction, which is a junction between two superconductors. These junctions are very attractive for information processing and storage [78]. The physical quantity of interest in the long Josephson junction is a quantum of magnetic flux, or a fluxon, which has a soliton behavior. The relevant nonlinear equation which describes the physical processes is the Sine–Gordon equation. This equation is particularly suitable in solid physics.

### Solitons in Optical Fibers

One of the fundamental applications of solitons concerns optical fibers. In optical fiber communication using linear pulses, dispersion and losses in the fiber limit the information capacity which can be transported and the distance of transmission. Hasegawa and Tappert [31] first proposed to balance the dispersive effect by nonlinearity using soliton-based communications. The Kerr effect, in other words the nonlinear change of the refractive index of the fiber material, is used and an initial optical pulse may form a nonlinear stable pulse, an optical soliton, which is in fact an envelope soliton. Hasegawa and Tappert proposed the previ-

ously mentioned nonlinear Schrödinger equation for the description of solitons propagating in optical fibers. The first observations of these solitons were made in 1980 by Mollenauer, Stolen, and Gordon [58]. In real media, the light intensity of the soliton decreases due to losses in the fiber, and suitable reshaping of the pulse is required. Numerous theoretical and experimental studies on nonlinear guided waves have been carried out, because of their wide range of applications [30,43]. Inserting an optical fiber in a laser cavity, a “soliton laser” may be obtained, a femtosecond laser which is now in use. A real revolution in international communications is taking place at present, based on optical soliton technology. It is anecdotal that a fiber-optic cable linking Edinburgh and Glasgow now runs beneath the path from which John Scott Russell made his initial observations, and along the aqueduct which now bears his name.

The NLS equation may be obtained from the Sine–Gordon equation. Previously considered systems as water waves on the free surface or chains of pendulums have also weak amplitude solutions, plane waves which have the form:

$$\theta = \psi e^{i(qx - \omega t)} + \text{c.c.} \quad (23)$$

where  $\psi$  is the wave amplitude,  $q$  is the wave number,  $\omega$  is the pulsation and c.c. denotes the complex conjugate of a complex number. When the wave amplitude is sufficiently increased for the nonlinearity to take place, modulation can spontaneously arise because harmonics are generated by the nonlinearity. The initial wave may split in waves packets whose properties correspond to the ones of solitons. These solitons consisting of a carrier wave which is modulated by an envelope signal are called envelope solitons. It can be shown [63] that the evolution of the envelope  $\psi$  is described by the NLS equation:

$$i \frac{\partial \psi}{\partial t} + P \frac{\partial^2 \psi}{\partial x^2} + Q |\psi|^2 \psi = 0 \quad (24)$$

where  $P$  and  $Q$  are coefficients which depend on the physical problem, as the significance of the variables  $t$  and  $x$ . The NLS equation is formally similar to the Schrödinger equation of quantum mechanics:

$$i \hbar \frac{\partial \psi}{\partial t} + \frac{\hbar^2}{2m} \frac{\partial^2 \psi}{\partial x^2} - U \psi = 0. \quad (25)$$

The potential  $U$  in Eq. (25) is proportional to the absolute square of the wave envelope  $\psi$ ,  $m$  is the mass of the quasi-particle and  $\hbar = h/2\pi$  where  $h$  is the Planck constant. The potential represents the self-trapping of the wave energy.

### Solitons in Solid Physics

Solitons are involved in the atomic structure of matter. At the beginning of the twentieth century, a central question in solid state physics was how a plastic deformation of a metal takes place in the corresponding crystal lattice. This problem induced numerous discussions, and it was suggested that if one atom in the lattice is pulled away, a neighbor atom follows the one which had been moved, jumping into the new hole, that is the new free lattice space. Then, a dislocation runs through the crystal. This proposition was improved in 1939 by Frenkel and Kontorova [22] who developed a model where not only each atom moves alone over one lattice distance, but also many atoms form a dislocation line generating a wave of translation. The process of traveling of the dislocation line through the crystal was assumed to occur without losses of energy. This is equivalent to Russell's solitary wave of translation which travels over long distance without change of form and velocity. This led Frenkel and Kontorova to make an analogy between the dislocation line and the soliton, and they obtained the one-soliton solution of the Sine–Gordon equation.

The solitons permit us to interpret properties of dielectric materials. In other respects, magnetic materials are interesting examples to experimentally verify in a very accurate way the theory of solitons at the atomic scale [57]. The concept of soliton in polymer physics constitutes a very nice case of an interdisciplinary approach. Their occurrence was suggested in 1988 by theoretician physicists [34]. Many studies have since then been carried out, in chemistry and experimental physics. Alan J. Heeger, Alan G. MacDiarmid and Hideki Shirakawa received the Nobel prize in chemistry in October 2000 for their works on the electric conduction of conductive plastics which is performed by solitons.

### Solitons in Biology

Interest in the physical modeling of biological processes has grown significantly since the beginning of the 1990s. Nonlinear localization phenomena have been recently shown in systems very close to biological systems [15]. The notion of the soliton is now used to explain the dynamics of biological macromolecules such as proteins and the molecule ADN; an approach to the dynamics of the molecule ADN can be obtained using the envelope solitons solution of the NLS equation [63]. Intense researches are presently being carried out on this subject. The great size of biological molecules allows collective behaviors of atom groups, which are with nonlinearity important components for the existence of solitons.

### Mathematical Methods Suitable for the Study of Solitons

Solitons may appear in many fields, such as fluid mechanics, solid state physics, plasma physics, optical fibers and biology, as shown in Sect. “Physical Properties of Solitons and Associated Applications”. The description of these physical systems often leads to dispersive and nonlinear equations which are not presently solvable. It is then important to refer to the great classes of soliton equations, which are idealized models for numerous systems. These great classes of soliton equations correspond to the Korteweg–de Vries equation, known as the KdV equation (Eq. (9)), the Sine–Gordon equation (SG; Eq. (22)), and the nonlinear Schrödinger equation (NLS, Eq. (24)). These three famous equations can be totally integrated. They are not completely dissociated. For instance, it is possible to obtain the NLS equation from the SG equation and from the KdV equation. The soliton solutions represent only a first approximation of the physical properties of real systems. Dissipation or a weak spatial or temporal variation of the physical parameters in real physical systems are examples of phenomena which are very common and which are not taken into account in the soliton equations. The modeling has to be adapted to the physical system and to the excitation conditions of this system. In the case of plasmas, the KdV equation is more suitable for a description of the system when intense pulses excite it, and the NLS equation is more adapted for sinusoidal perturbations.

There are two broad theoretical approaches available to obtain the solutions of the great classes of soliton equations, the analytic and the algebraic methods. Analytic approaches include the powerful inverse scattering transform (IST) [24] and the remarkable Hirota method [38]. One-soliton solutions or multi-soliton solutions can be obtained. Moreover, the IST method can predict the number of solitons that can emerge from an initial disturbance applied to a physical system [59]. This method shows how the solitons have a role in nonlinear normal modes as the Fourier modes for a linear equation [44]. There are also the Bäcklund transformation method [67], the direct linearizing transform method [1], the Painlevé analysis [45], and the method of position-like solutions [40]. The algebraic methods involve the Lie group theoretic method [60], the direct algebraic method [53], and the tangent hyperbolic method [51].

The theoretical methods cannot always be used since they need some conditions to make them applicable. Numerical methods can be applied [36]. Zabusky and Kruskal [82] were the pioneers in studying in 1965 the KdV equation numerically, by using the leapfrog method

as an explicit finite difference scheme. Many methods have since then been used: a split-step Fourier method [77], the hopscotch method [27], a pseudo-spectral method [21], spectral methods (the Galerkin and Chebyshev methods) [35,72], the finite element method [75], and Fourier collocation calculations [9].

### Future Directions

Solitons may occur in the macroscopic and microscopic scales of physics. This results from their very general existence conditions, the coexistence of dispersion and non-linearity. The concept of solitons is not restricted to a specific field of physics, and the investigation of solitons remains a very active research area. We have previously mentioned (Sect. “Physical Properties of Solitons and Associated Applications”) that one important application of solitons concerns optic fibers. The propagation of optical solitons through nonlinear fibers has been extensively studied; however, in real media, the dynamics of these solitons and the conditions for their generation are significantly affected by various inhomogeneities in the media. The problem of nonlinear wave propagation in the form of solitons in inhomogeneous optical fiber media is actually not well understood, despite the wide range of applications [30]. The investigation of nonlinear processes in coupled optical waveguides is another research direction for the future design of optical computers and sensor elements. Recent advances have shown that solitons have a great potential for the improvement of optical systems which demand fast and reliable data transfer. In the hydrodynamic field, the tsunamis often behave like solitary waves. The dramatic tsunami which occurred in the Indian Ocean on 26 December 2004 has clearly shown that the actual numerical models do not accurately predict the water propagation on the coastal plains. An important direction for future research is the (3D) three-dimensional simulation of tsunami breaking along a coast. The 3D simulation of the interaction of a tsunami with different beach profiles, with and without obstacles has also to be considered. This will permit the design of protecting devices and the setting up of security zones. In biology, nonlinear localization phenomena have been proved in model systems close to biological systems, but the challenge is open for biological molecules. Other important application fields for solitons concern magneto-electronics and secure communications. The formation of spatio-temporal patterns on perturbations of soliton systems are important problems to be studied in the years to come. The existence of instabilities in the solitary wave solutions of the great classes of soliton equations have to be tackled. Nonlinear excita-

tions in systems whose spatial dimension is greater than one raise many questions. Among these questions, there are the problems of the possible stable structures, their collision properties and the effect of external forces. New applications will certainly emerge from these studies.

### Bibliography

#### Primary Literature

1. Ablowitz MJ, Clarkson PA (1991) Solitons, nonlinear evolutions equations and inverse scattering. Cambridge University Press, Cambridge
2. Airy GB (1845) Tides and waves. *Encycl Metropolitana* 5:291–396
3. Bazin H (1865) Recherches expérimentales relatives aux remous et à la propagation des ondes. In: Darcy H and Bazin H (eds) *Recherches hydrauliques*, imprimerie impériale, Paris
4. Benjamin TB, Feir JE (1967) The disintegration of wavetrains on deep water. Part 1 Theory. *J Fluid Mech* 27:417–430
5. Boussinesq J (1871) Théorie de l'intumescence liquide appelée onde solitaire ou de translation, se propageant dans un canal rectangulaire. *CR Acad Sci* 72:755–759
6. Boussinesq J (1877) Essai sur la théorie des eaux courantes. *MSE* 23:1–680
7. Brandt P, Rubino A, Alpers W, Backhaus JO (1997) Internal waves in the Strait of Messina studied by a numerical model and synthetic aperture radar images from the ERS 1/2 satellites. *J Phys Oceanogr* 27:648–663
8. Bullough RK (1988) The Wave Par Excellence, the solitary progressive great wave of equilibrium of the fluid: an early history of the solitary wave. In: Lakshmanan M (ed) *Solitons: Introduction and applications*. Springer Ser Nonlinear Dyn. Springer, New York, pp 150–281
9. Canuto C, Hussaini MY, Quarteroni A, Zang TA (1988) *Spectral methods in fluid dynamics*. Springer, Berlin
10. Chanson H (2005) Le tsunami du 26 décembre 2004: un phénomène hydraulique d'ampleur internationale. *Premiers constats. Houille Blanche* 2:25–32
11. Cooker MJ, Weidman PD, Bale DS (1997) Reflection of high amplitude wave at a vertical wall. *J Fluid Mech* 342:141–158
12. Darrigol O (2003) The spirited horse, the engineer, and the mathematician: water waves in nineteenth-century hydrodynamics. *Arch Hist Exact Sci* 58:21–95
13. Dias F, Dutykh D (2007) Dynamics of tsunami waves. *Extreme Man-Made and Natural Hazards in Dynamics of Structures*. In: Ibrahimbegovic A, Kozar I (eds) *Proc NATO Adv Res Workshop on Extreme Man-Made and Natural Hazards in Dynamics of Structures*. Springer, Opatija, Croatia
14. Donnelly C, Chanson H (2002) Environmental impact of a tidal bore on tropical rivers. In: *Proc 5th Int River Management Symp*. Brisbane, Australia
15. Edler J, Hamm P (2002) Self-trapping of the amide I band in a peptide model crystal. *J Chem Phys* 117:2415–2424
16. Emerson GS (1977) John Scott Russell: a great Victorian engineer and naval architect. John Murray, London
17. Ezersky AB, Polukhina OE, Brossard J, Marin F, Mutabazi I (2006) Spatiotemporal properties of solitons excited on the surface of shallow water in a hydrodynamic resonator. *Phys Fluids* 18:067104

18. Fermi E, Pasta J, Ulam S (1955) Studies of nonlinear problems. Los Alamos report, LA-1940. published later In: Segré E (ed)(1965) *Collected Papers of Enrico Fermi*. University of Chicago Press
19. Fochesato C, Grilli S, Dias F (2007) Numerical modelling of extreme rogue waves generated by directional energy focusing. *Wave Motion* 44:395–416
20. Ford JJ (1961) Equipartition of energy for nonlinear systems. *J Math Phys* 2:387–393
21. Fornberg B, Whitham GB (1978) A numerical and theoretical study of certain nonlinear wave phenomena. *Philos Trans R Soc London* 289:373–404
22. Frenkel J, Kontorova T (1939) On the theory of plastic deformation and twinning. *J Phys* 1:137–149
23. Gardner CS, Morikawa GK (1960) Similarity in the asymptotic behaviour of collision-free hydromagnetic waves and water waves. Technical Report NYO-9082, Courant Institute of Mathematical Sciences. New York University, New York
24. Gardner CS, Green JM, Kruskal MD, Miura RM (1967) Method for solving the Korteweg-de Vries equation. *Phys Rev Lett* 19:1095–1097
25. Gerkema T, Zimmerman JTF (1994) Generation of nonlinear internal tides and solitary waves. *J Phys Oceanogr* 25:1081–1094
26. Goring DG (1978) Tsunamis – The propagation of long waves onto a shelf. Ph D thesis. California Inst Techn, Pasadena, California
27. Greig IS, Morris JL (1976) A hopscotch method for the KdV equation. *J Comput Phys* 20:64–80
28. Guizien K, Barthélemy E (2002) Accuracy of solitary wave generation by a piston wave maker. *J Hydraulic Res* 40(3):321–331
29. Hammack JL, Segur H (1974) The Korteweg-de Vries equation and water waves. Part 2. Comparisons with experiments. *J Fluid Mech* 65:289–314
30. Hao R, Li L, Li ZH, Xue W, Zhou GS (2004) A new approach to exact soliton solutions and soliton interaction for the nonlinear Schrödinger equation with variable coefficients. *Opt Commun* 236:79–86
31. Hasegawa A, Tappert F (1973) Transmission of stationary nonlinear optical pulses in dispersive dielectric fiber: II. Normal dispersion. *Appl Phys Lett* 23:171–172
32. Hashizume Y (1985) Nonlinear pressure waves in a fluid-filled elastic tube. *J Phys Soc Japan* 54:3305–3312
33. Hashizume Y (1988) Nonlinear pressure wave propagation in arteries. *J Phys Soc Japan* 57:4160–4168
34. Heeger AJ, Kivelson S, Schrieffer JR, Su WP (1988) Solitons in conducting polymers. *Rev Modern Phys* 60:781–850
35. Helal MA (2001) Chebyshev spectral method for solving KdV equation with hydrodynamical application. *Chaos Solit Fractals* 12:943–950
36. Helal MA (2002) Review: Soliton solution of some nonlinear partial differential equations and its applications in fluid mechanics. *Chaos, Solitons and Fractals* 13:1917–1929
37. Helal MA, El-Eissa HN (1996) Shallow water waves and KdV equation (oceanographic application). *PUMA* 7(3–4):263–282
38. Hirota R (1971) Exact solution of the Korteweg-de Vries equation for multiple collisions of solitons. *Phys Rev Lett* 27:1192–1194
39. Jackson EA (1963) Nonlinear coupled oscillators. I. Perturbation theory: ergodic problems. *J Math Phys* 4:551–558
40. Jaworski M, Zagrodzinski J (1995) Position and position-like solution of KdV and Sine-Gordon equations. *Chaos Solit Fractals* 5(12):2229–2233
41. Kharif C, Pelinovsky E (2003) Physical mechanisms of the rogue wave phenomenon. *Eur J Mech B/Fluids* 22:603–634
42. Korteweg DJ, De Vries G (1895) On the change of form of long waves advancing in a rectangular channel, and on a new type of long stationary waves. *Phil Mag* 39(5):442–443
43. Kruglov VI, Peacock AC, Harvey JD (2003) Exact Self-Similar Solutions of the Generalized Nonlinear Schrödinger Equation with Distributed Coefficients. *Phys Rev Lett* 90(11):113902 <http://prola.aps.org/abstract/PRL/v90/i11/e113902>
44. Lakshmanan M (1997) Nonlinear physics: integrability, chaos and beyond. *J Franklin Inst* 334B(5/6):909–969
45. Lakshmanan M, Sahadevan R (1993) Painlevé analysis, Lie symmetries and integrability of coupled nonlinear oscillators of polynomial type. *Phys Rep* 224:1–93
46. Lamb H (1879) *Treatise on the motion of fluids*. Hydrodynamics, 6th edn 1952. Cambridge University Press, Cambridge
47. Lamb H (1971) Analytical descriptions of ultrashort optical pulse propagation in a resonant medium. *Rev Mod Phys* 43:99–124
48. Liu PL-F, Synolakis CE, Yeh HH (1991) Report on the International Workshop on long-wave run-up. *J Fluid Mech* 229:675–688
49. Lo EYM, Shao S (2002) Simulation of near-shore solitary wave mechanisms by an incompressible SPH method. *Appl Ocean Res* 24:275–286
50. Lynch DK (1982) Tidal bores. *Sci Am* 247:134–143
51. Malfluet W (1992) Solitary wave solutions of nonlinear wave equations. *Am J Phys* 60(7):650–654
52. Marin F, Abcha N, Brossard J, Ezersky AB (2005) Laboratory study of sand bedforms induced by solitary waves in shallow water. *J Geophys Res* 110(F4):F04S17
53. Martinez Alonso L, Olmedilla Morino E (1995) Algebraic geometry and soliton dynamics. *Chaos Solit Fractals* 5(12):2213–2227
54. Maxworthy T (1976) Experiments on collision between solitary waves. *J Fluid Mech* 76:177–185
55. Mei CC, Li Y (2004) Evolution of solitons over a randomly rough seabed. *Phys Rev E* 70:016302
56. Michallet H, Barthélemy E (1998) Experimental study of interfacial solitary waves. *J Fluid Mech* 366:159–177
57. Mireska HJ, Steiner M (1991) Solitary excitations in one-dimensional magnets. *Adv Phys* 40:196–356
58. Mollenauer LF, Stolen RH, Gordon JP (1980) Experimental observation of picosecond pulse narrowing and solitons in optical fibers. *Phys Rev Lett* 45:1095–1098
59. Newell AC (1985) *Solitons in mathematics and physics*. SIAM, Philadelphia
60. Olver PJ (1986) *Applications of Lie groups to differential equations*. Graduate Texts in Mathematics, vol 107. Springer, Berlin
61. Ostrovsky LA, Stepanyants YA (1989) Do internal solitons exist in the ocean? *Rev Geophys* 27:293–310
62. Paquerot JF, Remoissenet M (1994) Dynamics of nonlinear blood pressure waves in large arteries. *Phys Lett A* 194:77–82
63. Peyrard M, Dauxois T (2004) *Physique des solitons*. EDP Sciences. CNRS Editions, Paris
64. Rayleigh L (1876) On waves. *Phil Mag* 5(1):257–279
65. Remoissenet M (1999) *Waves called solitons – Concepts and experiments*. Springer, Berlin



66. Rodwell MJ, Allen ST, Yu RY, Case MG, Bhattacharya U, Reddy M, Carman E, Kamegawa M, Konishi Y, Puls J, Pullela R (1994) Active and nonlinear wave propagation devices in ultrafast electronics and optoelectronics. *Proc IEEE* 82:1037–1058
67. Rogers C, Shadwick WF (1982) *Bäcklund transformations and applications*. Academic, New York
68. Rottman JW, Grimshaw R (2001) Atmospheric internal solitary waves. In: *Environmental stratified flows*. Kluwer, Boston, pp 61–88
69. Russell JS (1837) Report on the committee on waves. In: Murray J (ed) *Bristol, Brit Ass Rep*, London, pp 417–496
70. Russell JS (1839) Experimental researches into the laws of certain hydrodynamical phenomena that accompany the motion of floating bodies, and have not previously been reduced into conformity with the known laws of the resistance of fluids. *Trans Royal Soc Edinb* 14:47–109
71. Russell JS (1844) Report on waves. In: Murray J (ed) *Brit Ass Rep Adv Sci* 14, London, pp 311–390
72. Sanz-Serna JM, Christie I (1981) Petrov-Galerkin method for nonlinear dispersive waves. *J Comput Phys* 39:94–102
73. Stokes GS (1847) On the theory of oscillatory waves. *Trans Cambridge Phil Soc* 8:441–473
74. Su CH, Gardner CS (1969) Korteweg-de Vries equation and generalizations. III. Derivation of the Korteweg-de Vries equation and Burgers equation. *J Math Phys* 10:536–539
75. Taha TR, Ablowitz MJ (1984) Analytical and numerical aspects of certain nonlinear evolution equations, (III) numerical, KdV equation. *J Comput Phys* 55:231–253
76. Taniuti T, Wei CC (1968) *J Phys Soc Jpn* 24:941–946
77. Tappert F (1974) Numerical solution of the KdV equation and its generalisation by split-step Fourier method. *Lec Appl Math Am Math Soc* 15:215–216
78. Ustinov AV (1998) Solitons in Josephson junctions. *Phys D* 123:315–329
79. Weidman PD, Maxworthy T (1978) Experiments on strong interaction between solitary waves. *J Fluid Mech* 85:417–431
80. Yomosa S (1987) Solitary waves in large blood vessels. *J Phys Soc Japan* 56:506–520
81. Yuen HC, Lake BM (1975) Nonlinear deep water waves: theory and experiments. *Phys Fluids* 18:956–960
82. Zabusky NJ, Kruskal MD (1965) Interaction of solitons in a collisionless plasma and the recurrence of initial states. *Phys Rev Lett* 15:240–243
83. Zakharov VE, Shabat AB (1972) Exact theory of two-dimensional self-focusing and onedimensional self-modulation of waves in nonlinear media. *Sov Phys JETP* 34:62–69

## Books and Reviews

- Agrawal GP (2001) *Nonlinear Fiber Optics*. Academic Press, Elsevier
- Akmediev NN, Ankiewicz A (1997) *Solitons, Nonlinear Pulses and Beams*. Chapman and Hall, London
- Braun OM, Kivshar YS (2004) *The Frenkel–Kontorova Model. Concepts, Methods and Applications*. Springer, Berlin
- Bullough RK, Caudrey P (1980) *Solitons*. Springer, Heidelberg
- Davydov AS (1985) *Solitons in Molecular Systems*. Reidel, Dordrecht
- Dodd RK, Eilbeck JC, Gibbon JD, Morris HC (1982) *Solitons and Nonlinear Wave Equations*. Academic Press, London
- Drazin PG, Johnson RS (1993) *Solitons: an Introduction*. Cambridge University Press, Cambridge

- Eilenberger G (1981) *Solitons: Mathematical Methods for Physicists*. Springer, Berlin
- Hasegawa A (1989) *Optical Solitons in Fibers*. Springer, Heidelberg
- Infeld E, Rowlands G (2000) *Nonlinear waves, Solitons and Chaos*. Cambridge University Press, Cambridge
- Lamb GL (1980) *Elements of Soliton Theory*. Wiley, New York
- Toda M (1978) *Theory of Nonlinear Lattices*. Springer, Berlin

## Solitons Interactions

TARMO SOOMERE

Center for Nonlinear Studies, Institute of Cybernetics,  
Tallinn University of Technology, Tallinn, Estonia

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction: Key Equations, Milestones, and Methods](#)

[Extended Definitions](#)

[Elastic Interactions of One-Dimensional  
and Line Solitons](#)

[Geometry of Oblique Interactions of KP Line Solitons](#)

[Soliton Interactions in Laboratory and Nature](#)

[Effects in Higher Dimensions](#)

[Applications of Line Soliton Interactions](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Solitary wave** A localized structure which travels with constant speed and shape in otherwise homogeneous medium.

**Soliton** In the classical nomenclature, a soliton is a nonlinear spatially or temporally localized entity (solitary wave, impulse, wave packet, kink, etc.) of permanent form that retains its identity and shape in interactions with other similar entities. Also frequently used term for solutions of the relevant nonlinear partial differential equations approximately describing such physical entities.

**Soliton (colloquially as well as in certain domains)**

Used for denoting any long-lived localized structures (e.g. solitary waves, self-trapped optical beams, localized vortices) which exhibit low energy radiation and approximately keep their shape.

**Soliton solution** A solution, usually in closed form, of the nonlinear partial differential equations that exhibit properties of a single soliton. A multi-soliton solution exhibits properties and collective behavior of groups of solitons.



**Elastic interaction** A generalization of the concept of elastic collisions of ideal mechanical bodies: interaction of localized entities in which the dynamically significant properties of the counterparts such as the total energy, linear and angular momentum, and topological charge or spin are conserved. For low-dimensional and scalar solitons denotes the case in which all the properties of interacting solitons are completely restored after interaction.

**Inelastic interaction** Interaction of soliton-like structures that modifies one or more dynamically significant properties of the counterparts.

**Phase shift** A change of the position of a soliton owing to interactions with other solitons compared to the location in which it would be in the absence of other solitons.

**Resonant interaction** A specific form of soliton interactions leading to the fusion of two or more solitons into a new soliton. Generally associated with an infinite phase shift of the counterparts.

### Definition of the Subject

The concept of solitons reflects one of the most important developments in science of the 20th century: the nonlinear description of the world. It is customary to start an introduction to solitons by recalling that it is not easy to give a comprehensive and precise definition of a soliton. Frequently, a soliton is explained as a spatially localized (solitary) wave with spectacular stability properties. Although the combination of simultaneously being a localized structure propagating while mostly keeping its shape as waves do, and surviving for a long time in realistic (that is, nonlinear) conditions, already is a fascinating property in itself, yet another key quality defines whether a particular entity is a soliton. The distinction is made based on the way in which two or more of these objects interact with each other.

The classical nomenclature associates the term soliton with (i) nonlinear and (ii) localized entities, which (iii) have a permanent form and (iv) retain their identity in interactions with other entities from the same class (e.g. Drazin and Johnson [37]; also the entry ► [Solitons and Compactons](#)). The fundamental property of a soliton is thus to retain its identity in nonlinear interactions. In low-dimensional systems (understood here as environments admitting solitons in one spatial dimension and line solitons in two dimensions), a soliton's amplitude, shape, and velocity are restored after each interaction, whereas in more complex systems this request embraces all dynamically significant qualities. In other words, a structure

is a soliton if and only if its *interactions* are *fully elastic*. Thus, the interaction of solitons is always an intrinsic topic whenever solitons are considered.

In the nonlinear world, *linear superposition* does not hold and *nonlinear interaction* takes place. Mathematically, classes of entities are called linear, whenever any linear combination of entities belongs to the same class. Among structures corresponding to the solutions of certain equations, exclusively those that satisfy linear equations are denoted as linear ones. Any linear combination of solutions of a linear equation also solves this equation. The principle of linear superposition is that the resultant structure is simply the sum of the parties and it implies that the individual linear wave shapes are perfectly restored after meeting with each other.

The property of elasticity of soliton interactions has a fundamentally different nature, because, as a rule, a linear superposition of soliton solutions is not a solution of the governing nonlinear equation. This property distinguishes nonlinear *solitary* traveling waves (that is, spatially or temporally localized pulses, whereas the relevant single-wave solution of the underlying equation may be completely stable) from *solitons*, which may be observed and analyzed as separate entities but which always express property (iv) should similar entities appear in the system.

The propagation and interaction of both solitons and solitary waves in realistic conditions (that is, in the presence of dissipation, external fields, and inhomogeneities of the medium) frequently results in a certain loss of energy (usually in the form of radiation of relatively short waves); in certain cases collapse or fission of the structure, or a net exchange of energy between the waves occurs. Even if a solitary wave travels without radiation losses in an ideal environment, it may only be called a soliton if neither radiation loss nor net energy exchange occurs when it meets a similar structure; otherwise a perfect re-emerging of each soliton after the interaction would be impossible. More generally, no net changes of any dynamically or topologically significant quantities are permitted in elastic collisions. Only changes of certain dynamically less significant properties such as the exact location (phase) of the solitons are common in elastic interactions. Solitary waves or structures possessing a selection of properties of solitons are sometimes called quasi-solitons.

Although the first evidence of solitons was extracted from observations in nature, the theory of solitons and their interactions has been developed very much in terms of the mathematical theory of the relevant nonlinear partial differential equations (PDEs) of the motion. The definitions of solitons and their interactions are commonly formulated in terms of solutions to such equations. It is

even customary to speak of solitary waves and solitons, and of their interactions, having in mind the relevant exact solutions to these equations. Therefore, in mathematical terms, a soliton – either a solitary wave or a more complex entity – is a solution of a nonlinear PDE that elastically interacts with other similar solutions. In the contemporary nomenclature it is also customary to associate solitons and their interactions with numerical (multi-soliton) solutions to these equations.

Interactions between solitons are the most fascinating features of soliton phenomena. The most instructive quality of soliton interactions is the universality of many of their features that exists in spite of the extremely diverse physical origins of the counterparts. The presentation of the basic conceptual ideas involving soliton interactions, a discussion of several particular cases of such interactions and an overview of observations of the relevant phenomena in natural conditions are mostly limited to effects occurring in the elastic interactions of one-dimensional and line solitons, in which the solitons behave very much like ideal mechanical bodies and which serve as the kernel of the classical concept of solitons and their interactions. Some spectacular features of nearly elastic interactions in higher dimensions are interpreted in terms of generalizations of the classical elastic interactions. Finally, a selection of practical applications of specific features of soliton interactions is again discussed from the viewpoint of line solitons.

## Introduction: Key Equations, Milestones, and Methods

### Integrable Equations

The quality of some nonlinear PDEs to admit soliton solutions is associated with the property of integrability. Integrable equations usually admit an infinite number of integrals (constants) of motion. Many nonintegrable equations possess localized shape-preserving traveling wave solutions that resemble solitons. However, only integrable equations have the universal property of possessing exact multi-soliton solutions that reflect perfectly elastic interactions between individual solitons. Thus, the integrability of the underlying equation is a primary constituent of the classical soliton interactions. Arnold [10] defines, for example, a soliton as a solitary wave (solution) of an integrable PDE having additional stability and robustness features which are inherited directly from the integrability of this PDE. Nonlinear interactions even between the greatly different (but physically relevant) solutions of integrable equations are also elastic.

Among the variety of integrable PDEs that admit multi-soliton solutions, examples related to the Korteweg–de Vries (KdV) equation, the (focusing) nonlinear Schrödinger (NLS) equation, the integrable Boussinesq equation, the sine-Gordon (SG), and the Kadomtsev–Petviashvili (KP) equation will be used below. Below we shall refer to the listed equations as the classical soliton equations. Since the KdV and the KP equations represent a large number of different physical systems and their solutions can be easily visualized in terms of the easily accessible environment of shallow-water waves, solutions to and results from studies of these equations will be largely used for illustrations of the explanations below.

The rapid development of the theory of solitons has led to the discovery of many integrable equations which show multi-soliton solutions and describe soliton interactions (see, e.g. Arnold [10] and the entry ► [Partial Differential Equations that Lead to Solitons](#)). A number of analytical methods have been developed to obtain the (multi-) soliton solutions of integrable equations, starting from the late 1960s. A major tool is the *Inverse Scattering Transform* (IST), first developed by Gardner, Green, Kruskal and Miura [47,50] for the KdV equation. The method consists of associating to the evolution equation a Sturm–Liouville problem. In the case of the KdV equation the Sturm–Liouville equation is just the time independent Schrödinger equation of quantum mechanics, where the potential is the wave function of the KdV equation. The single- and multi-soliton solutions are associated with the discrete spectrum of eigenvalues. A generalization of the Inverse Scattering Transform, known as the *AKNS method* (named after Ablowitz, Kaup, Newell and Segur [2] who developed the method), applies to a large class of integral equations. See the entry ► [Inverse Scattering Transform and the Theory of Solitons](#) for detailed information.

At the same time, Hirota [62] found that soliton solutions can be obtained by reducing, through a suitable transformation, the evolution equation to a *bilinear form*. Using some properties of the bilinear operator, it is possible to find the multi-soliton solution of an integrable equation. The proof of integrability of a particular PDE is usually nontrivial. The experience with the Hirota method, applied to a variety of nonlinear wave equations, is that exact multi-soliton solutions are found in many cases when it is not known if the equation is integrable. It is natural to expect that the existence of such solutions reflects the integrability in some sense; however, no proof of this conjecture is known at this time.

Mathematically speaking, it is always possible to build an integrable equation representing certain features of an

observed phenomenon. Nevertheless, only a few of these equations can be properly derived, normally using asymptotic expansions, from the corresponding primitive equations describing physical phenomena. Even the classical soliton equations only approximately describe the natural phenomena. A more exact description of the natural processes requires the introduction of additional contributions to these equations. Such contributions usually destroy the property of integrability and lead to nonelasticity of interactions of soliton-like entities. Consequently, most of the times solitons are just a good approximation of solitary waves in nature. This limitation is to some extent balanced by the fact that in many cases integrable soliton-admitting equations adequately represent the basic features of physical phenomena even far beyond their formal scope of validity. In spite of such intrinsic limitations, the tools developed in soliton theory have allowed researchers to reach a very deep understanding of some physical phenomena which would have hardly been explained by other means.

### Milestones

Several milestones of the solitons' history (► [Solitons: Historical and Physical Introduction](#)) are particularly significant to soliton interactions. Already the famous first observations in 1834 of shallow-water solitons [132] implicitly involved certain effects created through soliton interactions. The observed wave ahead of a ship probably had a straight crest, as the one reproduced in 1995 in the Union Canal near Edinburgh (see, e. g. p. 11 in [34]). It is a remarkable feature of ship-induced solitons in relatively narrow and shallow channels that a straight-crested upstream soliton exists ahead of a two-dimensional (2D) disturbance, whereas the solitons generated ahead of a ship sailing in wider shallow water areas have curved crests. The straight crest of the Russell's soliton and of her sister structures are thus not intrinsically one-dimensional (1D) structured solitons. They are formed owing to a specific mechanism of soliton interactions as described below. The KP equation, allowing a simple but adequate description of the crest-straightening phenomenon, was derived even later than the word soliton was coined.

A substantial step towards the concept of solitons was made in the mid-1950s by Fermi, Pasta and Ulam [40] who numerically analyzed the evolution of phonons in an anharmonic lattice. From the mathematical point of view this problem can be described by a discretization of the KdV equation. Surprisingly at the time, the process did not lead to an energy equipartition among the modes, although nonlinearity generally tends to create such an equiparti-

tion. Instead, a sort of recurrence, an early evidence of the elastic soliton interactions, was observed.

The basic features of soliton interactions were first demonstrated for media described by the KdV equation [64,174]. The evolution of a sinusoidal initial wave with periodic boundary conditions gives rise to an ensemble of solitary waves that move with different speeds and gradually overtake each other. Originally eight of them were mentioned; later revisitations increased the count to nine [106]. The difference in the number of solitons is not principal since small-amplitude solitons only become evident through additional phase shifts in extremely long-term simulations [39,133]. One surprise of the system was that after a very long time the whole profile reappears. The other, conceptually much more striking effect, was the persistence of the waves: after a certain phase of interaction with each other, they continued thereafter as if there had been no interaction at all. This persistence, which reflects the essence of the soliton interactions, led Zabusky and Kruskal to invent the name "soliton".

Solitons and new features of their interactions were later discovered in a large number of different physical systems. The universal principle generalizing the physical forces affecting the birth of all solitons is that the propagating disturbance (a pulse in a fiber, a wave-packet on a water surface, a self-focusing optical beam, a localized vortex, or a vortex dipole, etc.) is captured in a potential well jointly induced by its motion and by virtue of the nonlinearity of the underlying system. The solitons can thus be interpreted as the bound states of the relevant potential wells, and soliton interactions – as interactions between the bound states of a jointly induced potential well, or between bound states of different wells located at close proximity. This concept of potential well becomes vividly evident in the inverse scattering theory due to the conservation of the eigenvalues of the underlying spectral problem. This universality explains why, in spite of the immense diversity of the solitons and the underlying physical systems, the interactions (collisions) between solitons in all of these systems follow the same principles.

### Classical Soliton-Admitting Equations and Appearance of Solitons

The particular appearance of soliton interactions may vary a lot, depending on the physical system and the nature of the solitons. The Russell soliton is an example of nontopological solitons that often become evident in the form of traveling waves and that cannot exist in rest. On the contrary, single topological solitons (that interpolate between

two different states of the system which have the same energy) may exist in rest.

A localized entity consists of an infinite number of Fourier harmonics and generally experiences (linear) spreading or wave radiation. A localized entity of permanent form therefore may only appear if the spreading effects are balanced by a certain restoring force that evidently can only be of a nonlinear nature. In other words, all solitons correspond to a certain balance between spreading and nonlinear effects, the latter becoming evident, for example, as nonlinear steepening of the wave shape in the KdV system, or focusing of diffractive structure in nonlinear optics.

The physical meaning of spreading depends on the particular system. Dispersion is responsible for the spreading of pulses in all media in which the group velocity depends on the wave properties. The classical examples of dispersive solitons are surface [132], internal [33] and Rossby solitons [128] in geophysical systems. Similar structures occur in a large variety of environments such as drift solitons or ion-acoustic solitons in plasma [169], or solitons in optical fibers [60].

Many examples of dispersive solitons and their interactions occur in the framework of the KdV equation and its generalizations. It is a characteristic equation governing weakly nonlinear, spatially one-dimensional (1D) waves whose phase speed attains a simple maximum for infinitely long waves

$$\tilde{\eta}_t + \frac{3c_0}{2h} \tilde{\eta} \tilde{\eta}_x + \frac{c_0 h^2}{6} \tilde{\eta}_{xxx} = 0, \quad (1)$$

in nondimensional variables  $\eta_t + 6\eta\eta_x + \eta_{xxx} = 0$ ,

where  $\eta$  is the relative water surface elevation in the shallow water environment (Fig. 1),  $h$  is the still water depth, and  $c_0$  is the maximum phase speed (celerity) of linear waves at this depth. Its generalization to the case of solitons propagating in slightly different directions

$$(\eta_t + 6\eta\eta_x + \eta_{xxx})_x + 3\eta_{yy} = 0, \quad (2)$$

was derived by Kadomtsev and Petviashvili [66]. Named the KP equation after them, it describes features of a so-called weakly 2D environment (also called 1.5-dimensional systems). The nondimensional  $(x, y, t, \eta)$  and physical variables  $(\tilde{x}, \tilde{y}, \tilde{t}, \tilde{\eta})$  in Eq. (2) are related as follows:  $x = \sqrt{\varepsilon}(\tilde{x} - \tilde{t}\sqrt{gh})/h$ ,  $y = \varepsilon\tilde{y}/h$ ,  $t = \sqrt{\varepsilon^3 gh}\tilde{t}/h$ ,  $\eta = 3\tilde{\eta}/(2\varepsilon h) + O(\varepsilon)$  whereas  $\varepsilon = |\tilde{\eta}_{\max}|/h \ll 1$ .

Since the KdV equation only admits nontopological elevation solitons, the KdV/KP framework misses several basic features of soliton interactions, such as collisions of

solitons with different topological charge and interactions of solitons of different type. These classes of dispersive solitons are represented by the sine-Gordon (SG) equation

$$\theta_{tt} - c_0^2 \theta_{xx} + \omega_0^2 \sin \theta = 0, \quad (3)$$

where  $\theta$  in many applications has the meaning of a certain  $2\pi$ -periodic angle and  $\omega_0$  is the minimum angular frequency of linear waves. This equation, with its name stemming from a play of words regarding its form which is similar to the Klein-Gordon equation, arose already in the 19th century in studies of certain surfaces. It grew greatly in importance when it was realized that it led to kink and antikink solutions with the collisional properties of solitons [116]. Its major field of application is solid state physics, yet it also appears in a number of other physical environments such as the propagation of fluxons in Josephson junctions (a junction between two superconductors), the motion of rigid pendulums attached to a stretched wire, and dislocations in crystals. Its kink solutions (Fig. 1)

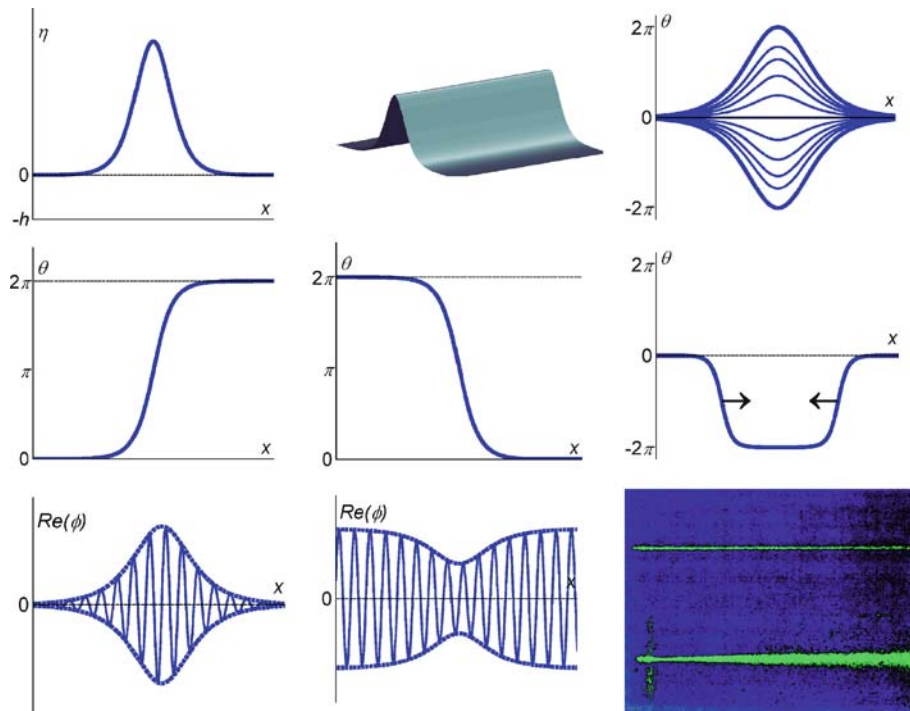
$$\theta_S = 4 \arctan \exp \left( \pm \frac{\omega_0}{c_0} \frac{x - vt - x_0}{\sqrt{1 - v^2/c_0^2}} \right), \quad (4)$$

where  $x_0$  is the initial position of the single soliton at the  $x$ -axis and  $v < c_0$  is the speed of translation of the soliton, represent so-called topological solitons that interpolate between two different states of the system, which have the same energy and may exist in rest (e.g. a frozen (anti)kink,  $v = 0$ ). The SG equation is a simple model admitting breather solitons and soliton pairs with different topological charges. The bulk topological charge is an invariant of the system. In a more complex manner the topological charge becomes evident in the theory of optical spatial solitons, where it can be interpreted in terms of spin of elementary particles. The kink and antikink solutions also represent the possibility of the existence of solitons and antisolitons. Both attractive and repulsive interactions can be vividly demonstrated in this framework and, after all, interactions of physically meaningful solitons of different types are possible.

Another key equation for dispersive waves is the nonlinear Schrödinger (NLS) equation that has a nondimensional form

$$i\psi_t + p\psi_{xx} + q|\psi|^2\psi = 0 \quad (5)$$

and only has soliton solutions in the focusing case  $pq > 0$ . This is a universal equation describing the evolution of complex wave envelopes  $\psi$  in a dispersive weakly nonlinear medium. It applies, among other phenomena, to



**Solitons Interactions, Figure 1**

A selection of solitons. From left to right: top – KdV traveling wave soliton, KP plane (line) soliton, SG breather; middle – kink, antikink, and a kink-antikink pair; bottom – bright and gray envelope soliton, and a top view photograph of a 10- $\mu\text{m}$ -wide optical spatial soliton propagating in a strontium barium niobate photorefractive crystal and the same beam diffracting naturally. The bottom right image is from [152]. Optical spatial solitons and their interactions: Universality and diversity. Reprinted with permission from AAAS

deep water waves. The studies of its soliton solutions and their interactions also date back to the 1960s [175,176]. Its major application in the context of soliton interactions is nonlinear optics, where its different versions describe the propagation and interactions of both dispersive solitons (e. g. envelope solitons in optical fibers) and diffractive optical solitons.

The basic reason for the existence of optical solitons is that the optical properties (refractive index or absorption) of some materials are modified by the presence of light. This feature introduces nonlinearity into the system and alters the propagation of optical pulses either in space or in time. A dispersive optical soliton is formed when the linear dispersion effects are balanced by a sort of lensing (of short light pulses of permanent form that are called temporal solitons in the optical nomenclature) in an appropriate material (for example, an optical fiber). They were predicted by Hasegawa and Tappert [60] and first observed by Mollenauer et al. [97]. The basic properties of their interactions [59] are analogous to those occurring for KdV or SG solitons. There is continuous interest for their studies because of their applications, e. g. in long-distance optical communication systems.

The generic example of diffractive solitons, evolution and interactions of a part of which are described by the NLS equation in two space dimensions, are optical spatial solitons. The natural tendency to broaden in space owing to the diffraction of optical beams (Fig. 1) can be balanced, for example, due to the optical Kerr effect that consists of a local refractive index change induced by light. As a first approximation, this change is assumed to take place instantaneously and to be proportional to the local intensity of light. In the focusing case (when an increase of the intensity of light causes an increase in the refractive index), the light-induced lensing can make an optical beam stable with respect to perturbations in both its width and intensity. The resulting beam in a 2D or 3D medium is an example of a diffractive soliton. Interaction of such beams occurs owing to overlapping of the modified regions of the medium. Usually such an interaction is local and has a long range only under specific conditions [129].

The first studies suggesting the existence of such solitons in nonlinear optical Kerr media date back to the 1960s [11,27]. In some cases (such as the 1D Kerr solitons in a planar medium) they are classical solitons and



described by a fully integrable equation (the NLS equation in two space dimensions [176]). Self-trapped light beams and more complex structures in a 2D and 3D medium attracted the attention of researchers starting from the 1980s, when the appropriate materials were found [13].

In multi-dimensional media, both diffraction and dispersion affect the propagation of solitary waves. A classical example of a diffractive-dispersive system is the motion of short-crested waves on the water surface. Most of the analysis of soliton interactions in this system has been made for effectively 1D solitons with infinitely long crests (so-called *plane* or *line solitons*); the effects of diffraction being implicitly neglected in their propagation and interactions. A practically realized example of diffractive-dispersive solitons are short pulses of incoherent (or white) light that is self-trapped both in the direction of its propagation and in the transversal direction(s), and that propagate changing neither their shape nor their length [94].

### Extended Definitions

The classical nomenclature of low-dimensional solitons and their interactions was formulated several decades ago, and it is not surprising that many later discoveries have led to attempts to extend its content. A part of the extensions add several consistent features (such as the conservation of linear and angular momentum, and topological charge or spin during the soliton interactions); however, in several schools substantial generalizations of the term soliton and of the interpretation of the soliton interactions have been introduced. The only component of the classical nomenclature kept in all the interpretations is the essential role of nonlinearity. Since it is virtually impossible to reflect all the extensions, mostly the material and results following the classical definition are presented below.

The solitons are commonly interpreted as a subset of traveling solitary waves. The classical nomenclature favors structures that are stationary in a proper coordinate system; yet it does not exclude the resting topological solitons (e. g. single-kink solutions to the sine-Gordon equation).

A principal extension of the classical definition of solitons that caused quite a serious discussion in the 1970s [101] consists of accounting for the *resonant interactions*. They may result in the fusion of several solitons into a new soliton or, equivalently, in the decay of solitons into counterparts [177]. Although the number of solitons is not conserved, all the parties of the resonant system are solitons at the limit of exact resonance and no energy radiation occurs. It is commonly accepted now as a variation (or a limiting case) of elastic interactions.

The classical soliton interactions preserve the shape of the solitons. The shape is understood in a wide sense; e. g. the shape of the envelope of the NLS solitons is preserved as well as the time-dependent shape of breathers that have an internal oscillation. This quality, which is universal for all scalar solitons, has a more general interpretation in the case of *vector* (composite) *solitons*. For example, the otherwise elastic interactions of Manakov vector solitons exhibit a shape-changing nature [125] in an integrable system consisting of two coupled NLS equations [82]. This property has a paramount practical importance and bright perspectives e. g. in optical soliton-based computations ([155], see also entry ► [Optical Computing](#)), being one of the key options of practical applications of specific features of soliton interactions.

An obvious reason for a wider view on soliton interactions is that the relatively simple integrable equations admitting soliton solutions stem from certain asymptotic expansions and only approximately describe the processes in nature. A more exact representation of the physical processes through inclusion of higher order terms of those expansions generally destroys the integrability of these equations. In many cases, though, the perturbing terms are small. The influence of a small dissipative perturbation is usually obvious: it brakes the moving solitons and/or damps the oscillating solitons. Such structures are sometimes called dissipative solitons although such a name is somewhat self-contradicting. Effects due to conservative (Hamiltonian) perturbations are much richer in content. Usually they do not destroy or brake solitons, but they may, e. g., render otherwise elastic soliton interactions to inelastic ones owing to extra wave radiation. A well-known example of the effect of perturbations is dissipation-induced annihilation of a kink-antikink pair in a perturbed sine-Gordon equation that otherwise would lead to a breather [114]. A monumental survey of the relevant problems is presented in [77]; see also entry ► [Soliton Perturbation](#).

Since none of the physical and numerical environments perfectly matches the governing equations, solitary waves both in realistic conditions and in numerical simulations always are approximations to solitons. In numerical experiments practically ideal solitons can be represented and errors basically occur due to purely computational inaccuracies. Laboratory experiments with solitons encounter problems with perfect excitation of even single solitons and moreover with controlling their interaction properties; however the relevant results are of fundamental importance in establishing the adequacy, the limits and the practical applicability of the properties of solitons and their interactions.

In the theory of elementary particles frequently a soliton is defined merely as a localized solution (resp. entity) of permanent form and the constraint of its survival in collisions is released (e. g. Rebbi [127]). A popular and deep example is the quantum field theory, where e. g. the Yang–Mills field equations admit solutions localized in space (which represent very heavy elementary particles), and also solutions, localized in time as well as in space (instantons). This position has certain support by the conjecture that the classical soliton equations can be interpreted as reductions of self-dual Yang–Mills equations and thus the classical solitons form a subset of the above solutions.

A similar position is also widely adopted in studies of optical beams and vortices. A large part of such beams are described by nonintegrable equations [135] and thus lie beyond the classical soliton nomenclature. Many features of their behavior and interactions, however, demonstrate a striking similarity with that of classical solitons, and exhibit universal properties [152], resembling the similar universality of the classical solitons interactions. Soliton-like beams organized by different nonlinear effects were mostly called solitary waves until the mid-1990s. The modern nonlinear optics nomenclature treats all self-trapped optical beams as solitons [136] although their interactions are not always elastic. A selection of results of studies of their interactions that give a flavor of the richness of soliton interaction phenomena in 3D media is presented below.

The constraint of energy conservation both in the motion and the interactions of the solitons is frequently disregarded in studies of long-lived structures which exhibit reasonable radiation [42] but that still survive in certain collisions. On the other hand, other highly interesting solitary structures such as (Rossby) dipole modons [83,84] do not radiate energy but only survive under certain conditions. A certain amount of results and concepts in this domain is presented to highlight the overlapping of the properties of soliton interactions and interactions of other long-living structures.

### Elastic Interactions of One-Dimensional and Line Solitons

The term soliton was originally introduced for nonlinear disturbances, the interaction of which resembles the collision of particles and is fully elastic (Table 1), so that the number of solitons is always conserved and their amplitudes are fully restored afterwards. In low-dimensional cases such as those described by the KdV equation, the amplitudes, directions and propagation velocities of each soli-

ton always recover their initial values (Fig. 2). The shape of each soliton is prescribed by the structure of (single-) soliton solutions to this equation and, as the simplest evidence of the recurrence of the system, it is not surprising that the shape is perfectly restored as well.

Three phases of the interactions of solitons can be distinguished either in time or in space (Fig. 2). First, well-defined, clearly separated solitons approach each other. In the second (interaction) phase, they usually lose their identity and merge into a composite structure. After a while, the solitons emerge again. The appearance of the composite structure depends on the particular physical system; as a rule it is neither a linear superposition of (properly shifted) counterparts nor a new soliton. Therefore solitons survive the collision event, even though they completely merge for a while. The variety of the composites range from the complete vanishing of the counterparts (e. g. kink-antikink collision in the SG equation) over certain interactions of KdV solitons, in which the individual identities are almost conserved and two humps are always visible, to a fourfold elevation of the water surface in a resonant interaction of the KP solitons.

In some environments, one cannot say whether the solitons propagate through each other as waves do or collide as particles do. Both interpretations have a solid ground. The interaction of unidirectional KdV solitons resembles the collision of two particles whereas, e. g. in head-on interactions of Rossby dipole solitons [84] the identity of both counterparts can be continuously tracked and the interaction process resembles a sort of complex overtaking.

### Attraction and Repulsion

Another feature of soliton interactions resembling collisions of real massive and/or electrically charged particles that exert certain forces on each other is that soliton interactions may show an *attractive* or *repulsive* nature. This feature can be demonstrated in the framework of the kink-antikink solution of the sine-Gordon equation

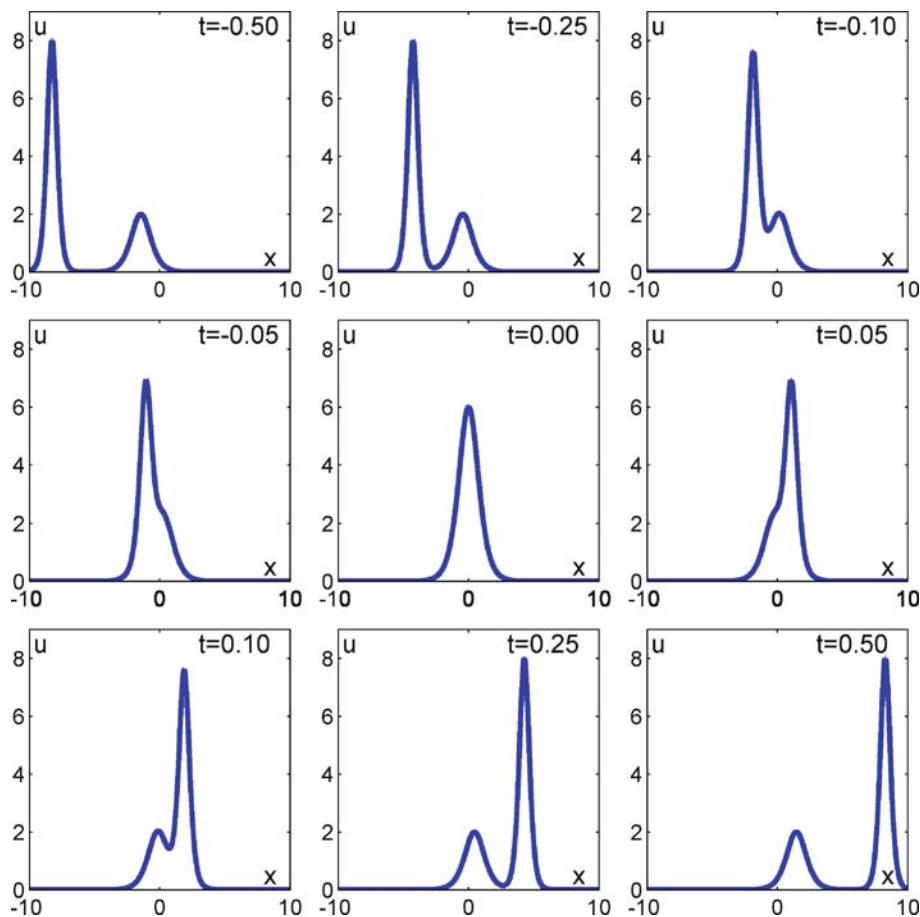
$$\theta_{SA} = 4 \arctan \frac{c_0 \sinh v t \xi}{v \cosh x \xi}, \quad \text{where } \xi = \frac{\omega_0}{c_0} \frac{1}{\sqrt{1 - v^2/c_0^2}} \quad (6)$$

and that only exists if  $v > 0$ . The motion of the counterparts speeds up when they move towards each other, and slows down when they move apart after passing each other. If one examines the positions of the centers of the counterparts of this solution, the sequences of their positions in the vicinity of their meeting point are not the same

Solitons Interactions, Table 1

Properties of solitons in elastic interactions of 1D and line solitons. Analogous rules hold for angular momentum and spin in elastic collisions in higher dimensions

| Property                               | 1D interactions and nonresonant interactions of line solitons                                                                                                                                                             | Exceptions in resonant interactions       |
|----------------------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------|
| Number of solitons                     | Conserved except in a short phase of interaction                                                                                                                                                                          | Changed                                   |
| Energy                                 | Restored after interactions for each soliton                                                                                                                                                                              | Merging possible; total energy conserved. |
| Amplitude Shape (steepness)            | Substantial local changes may occur in the interaction region; restored after interactions                                                                                                                                | Durable changes may occur                 |
| Phase or location                      | Finite phase shifts commonly occur; yet no phase shifts in certain environments. The total phase shift is exactly equal to the sum of partial shifts that would result from separate collisions with each of the solitons | Infinite phase shifts                     |
| Geometry                               | Substantial changes in the interaction region, restored after interaction                                                                                                                                                 | Durable changes may occur                 |
| Propagation direction, linear momentum | Conserved for each soliton                                                                                                                                                                                                | Conservation of bulk linear momentum      |
| Topological charge                     | Conserved for each soliton                                                                                                                                                                                                |                                           |



Solitons Interactions, Figure 2

Temporal evolution of KdV solitons described by the two-soliton solution  $u(x, t) = 12 \times (3 + 4 \cosh(2x - 8t) + \cosh(4x - 64t)) / ([3 \cosh(x - 28t) + \cosh(3x - 36t)]^2)$  [37]. The amplitudes of the counterparts are  $u_{1\max} = 8$  and  $u_{2\max} = 2$ . Notice the gradual decrease of the taller soliton when it catches the shorter, and its gradual increase when it moves further ahead of the shorter one, and the relatively small amplitude  $\dot{u} = 6$  of the composite structure

as they would be if they had been alone in the system, or far from the other counterpart.

If the solitons were massive or charged particles, one would say that an attractive force exists between them which decreases as their separation increases. On the contrary, the speed of two kinks as well as two antikinks approaching each other slows down as their separation decreases, which is characteristic for repulsive interaction. Another classical reflection of “forces” between solitons is the effect of attraction or repulsion of (envelope) solitons in optical fibers [53].

In two dimensions, the attractive interactions of two obliquely interacting line solitons become evident as an attraction of the relevant wave crests. In the case of optical spatial solitons the attraction affects their centroids. For repulsive interactions the opposite holds. The relevant effects are vividly demonstrated in the framework of the KP equation for the description of drift vortex solitons in environment, otherwise described by the Hasegawa–Mima equation [158]. This environment is equivalent to the one described by the Charney–Obukhov equation for large-scale motions in geophysical hydrodynamics admitting Rossby waves and solitons.

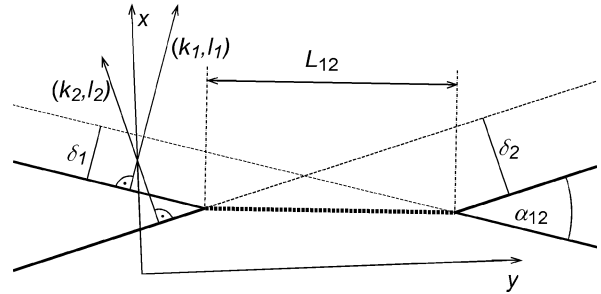
### Transient Amplitude Changes, Durable Phase Shifts and Recurrence Patterns

Already the first study of solitons [174] revealed some other universal aspects of soliton interactions. First, certain durable phase shifts may occur, the magnitude of which depends on the interaction details. For example, during the collision presented in Fig. 2, the taller soliton is shifted forward by an amount  $\Delta x_1 = \frac{1}{2} \log 3$  and the shorter one backwards by  $\Delta x_2 = \log 3$  compared to the case when the solitons are alone in the system. The above-discussed “forces”, if present in the system, may be interpreted as the cause of the phase shifts.

The total phase shift of a soliton induced by elastic collisions with any number of solitons is exactly equal to the sum of (partial) shifts that would result from separate collisions with each of the solitons involved. This property is commonly referred to as the *absence of many-particle effects*.

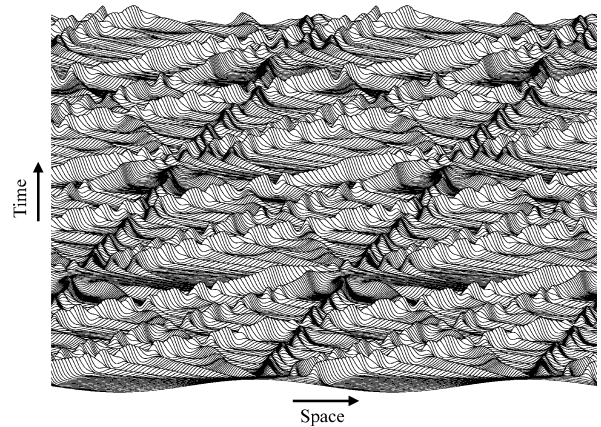
In the case of line solitons the phase shifts become evident in the form of durable shifts of the counterparts during the interaction (Fig. 3). In the negative phase shift case the solitons are delayed compared to their position in the absence of interactions, whereas in the positive phase shift case they seem to be accelerated.

Second, the amplitudes of the interacting solitons may gradually change as they approach to each other (Fig. 2).



**Solitons Interactions, Figure 3**

Idealized patterns of crests of interacting KP solitons (bold lines), their position in the absence of interaction (dashed lines) and the crest of the residue  $s_{12}$  (bold dashed line, see Eq. (7)) corresponding to the negative phase shift case (adapted from [120])



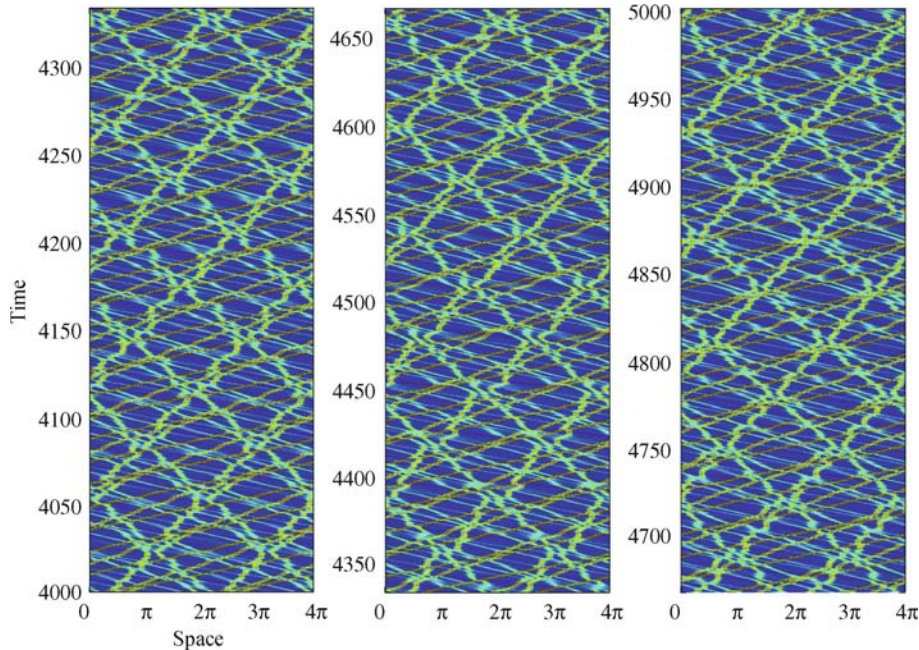
**Solitons Interactions, Figure 4**

Time-slice plot over two  $2\pi$  periods of the evolution of the ensemble of the KdV solitons generated from a sinusoidal initial condition in space for  $0 \leq t \leq 97$  at  $\log d = -2.3$ . Reprinted from [134]

The amplitude of the composite structure is generally not equal to the sum of amplitudes of the interacting solitons. In interactions of a small number of KdV solitons running along an infinite medium, the behavior of each counterpart and its influence on the partners can always be identified from the shape of the elevation since after some time the solitons become separated enough to detect all of them and their perfectly restored properties in the limit of infinite separation.

More insight into the properties of the soliton interaction is provided by long-term simulations of ensembles of solitons excited from a periodical signal (Fig. 4). The total number of interacting parties is infinite then, the solitons are tightly packed and such a separation not necessarily occurs. In some cases the presence of several shorter solitons can only be identified through their contributions





**Solitons Interactions, Figure 5**

Patterns of locations of crests of the KdV solitons generated from sinusoidal initial conditions at  $\log d = -2.3209$ . Reused with permission from [39]

to the phase shifts of the partners since they are seldom visually identifiable; such solitons are called hidden solitons [39].

The trajectories of the (crests of the) counterparts may show quite a complex pattern and may exhibit both simple recurrence (at which the initial system is approximately restored) as well as super-recurrence (understood as a nearly perfect restoration of the initial system over longer times).

Practically periodical in time rhombus-like patterns (suggesting that interactions of quite a small number of solitons govern the properties of the system) optionally occur in cases when a super-recurrence is possible (Fig. 5), whereas in other cases these patterns show gradual variation also over extremely long times. There seems to exist a critical value  $d^*$  in the range  $-1.8 < \log d^* < -1.9$  of the dispersion parameter  $d$  in the KdV equation (presented in the form  $\eta_t + \eta\eta_x + d\eta_{xxx} = 0$ ), which distinguishes if a super-recurrence exist. If  $d < d^*$ , no super-recurrence can be detected even within extremely long calculation times [39,134].

### Durable Local Amplitude Changes in Oblique Interactions of Line Solitons

The interaction of unidirectional KdV solitons does not create any large changes in the wave amplitudes ([37], see

also ► [Korteweg–de Vries Equation \(KdV\), History, Exact  \$N\$ -Soliton Solutions and Further Properties of the](#)). However, amplitude amplification may occur under certain conditions when line solitons propagating in slightly different directions meet each other. Extensive evidence comes from coastal engineering, where it was known already a long time ago that ocean waves approaching breakwaters and seawalls under a certain angle caused unexpectedly large overtopping. This phenomenon was systematically studied already in the 1950s [117], well before the word soliton was coined.

When a shallow-water wave is obliquely launched along an ideally reflecting wall, the reflection does not necessarily follow the rules of geometrical optics. For a certain range of incidence angles, the crests of the approaching and the reflected wave fuse together near the wall and the process resembles the *Mach reflection*. The common crest, an analogue of the *Mach stem*, is generally unsteady and gradually lengthens for sine waves [15] and Stokes waves [173]. This type of reflection also occurs during perfect reflection of random [89] and solitary waves of different nature [45,92,93,115,159].

For a certain set of parameters of the approaching KdV soliton, the resulting structure is equivalent to half of the pattern created by two interacting KP solitons of equal amplitude (Fig. 3). The prominent feature of both the (Mach)



reflection and oblique interactions of shallow-water solitons is that the height of the common crest may considerably exceed the sum of the heights of the interacting solitons. For interacting waves with equal amplitudes the hump can be up to four times as high as the incoming waves. Originally established in the context of the reflection of Boussinesq solitons ([92,93]; named Mach reflection by Melville 1980 [91]), the amplitude amplification in a certain region occurs in all oblique attractive interactions of line KP solitons (that is, in interactions with a negative phase shift). Since the interaction pattern is stationary in an appropriately moving coordinate system for a range of the parameters of the interacting solitons, it may lead to durable local changes of the amplitude of the resulting pattern of elevation.

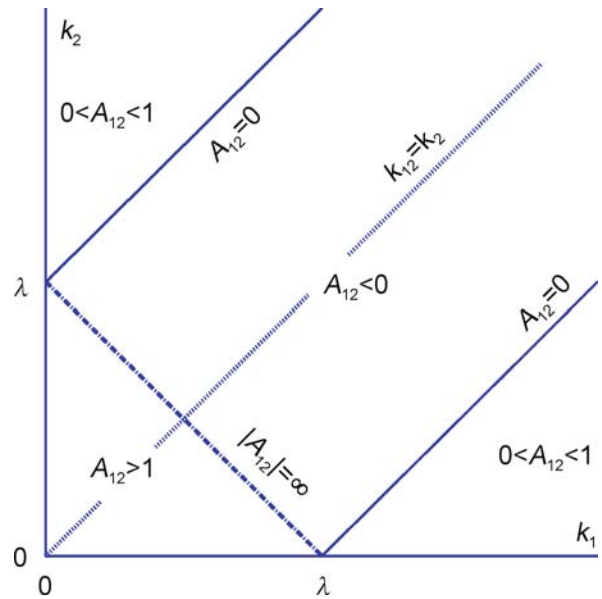
### Resonance

While the theory of integrable systems requires that the 1D solitons retain their identity in interactions, collisions of solitons in multiple dimensions may lead to the formation of new solitons. The phenomenon called *resonant interaction* leads to the emergence of new structures that combine the energy of the counterparts and to modifications of the number of solitons and of some of their geometrical features.

This possibility was first recognized by Newell and Redekopp [101] in the context of the KP and NLS equations. Miles [93] gives an early answer to their question about the practical consequences of this phenomenon in terms of an essential increase of the wave amplitude occurring during the resonant Mach reflection. Originally named “break-down of the Zakharov–Shabat theory of integrable systems with more than one spatial dimension”, it highlights the more complex nature of soliton interactions in more than one dimension, where large phase shifts are possible.

Remarkably, the resonance conditions  $\kappa_\infty = \kappa_1 + \kappa_2$ ,  $\omega_\infty = \omega_1 + \omega_2$ , where  $\kappa_i = (k_i, l_i)$ ,  $i = 1, 2, \infty$  are the wave vectors and  $\omega_i$  are the frequencies of the solutions of the corresponding linearized KP equation, are precisely the same as those for the resonant interaction of three weakly nonlinear waves. Conceptually, however, such interactions more resemble so-called double resonance, in which not only the above resonance conditions are met but also the group velocities of two or more counterparts are equal and otherwise sporadic energy exchange within the resonant triplets is replaced by much more intense interactions (e. g. [145]).

The patterns of line soliton crests (Fig. 3) suggest that only attractive interactions may result in resonance. The resonant interactions between line solitons correspond to



Solitons Interactions, Figure 6

Negative and positive phase shift areas for the two-soliton solution of the KP equation for fixed  $l_{1,2}$ . The phase shift  $\Delta_{12} = -\ln A_{12}$  has a discontinuity along the line  $k_1 + k_2 = \lambda$  at which  $|A_{12}| = \infty$  and the resonant interactions occur. No two-soliton solution exists in the area where  $A_{12} < 0$ . An analogue of linear superposition occurs when  $A_{12} = 0$ . Adapted from [120]

infinitely large phase shifts and infinitely strong attraction, and lead to durable changes of the geometry of the system. In the situation depicted in Fig. 3, the resonance would lead to an infinitely long common crest and thus render the “four-arm” pattern to a “three-arm” one.

In the KP framework, the resonance phenomenon is connected with the fact that its multi-soliton solutions representing the interaction of line solutions only exist for a subset of the space of parameters of the interacting solitons (Fig. 6). Resonance occurs when the parameters of the interacting solitons lie at a certain part of the border of this subset.

A new soliton born at the exact resonance may resonantly interact with yet another soliton. This mechanism has been employed in [17] to construct a family of exact solutions to the KP equation corresponding to resonance of several solitons. Such solutions consist of unequal numbers of incoming and outgoing line solitons, the parameters of which can be obtained from asymptotic analysis [16]. This class contains a variety of multisoliton solutions, many of which exhibit nontrivial spatial interaction patterns that are generally characteristic to multisoliton solutions to the KP equation [118].

## Geometry of Oblique Interactions of KP Line Solitons

### Patterns

A relatively simple but instructive, easily visualizable and rich-in-content model demonstrating the discussed features of interactions of line solitons is the KP equation. It admits simple explicit formulae for multi-soliton solutions and extensive analytical treatment of their geometrical properties. The two-soliton solution to the KP equation can be written as a sum  $\eta = s_1 + s_2 + s_{12}$  of terms, reflecting to some extent the two interacting solitons  $s_1, s_2$  and residue  $s_{12}$

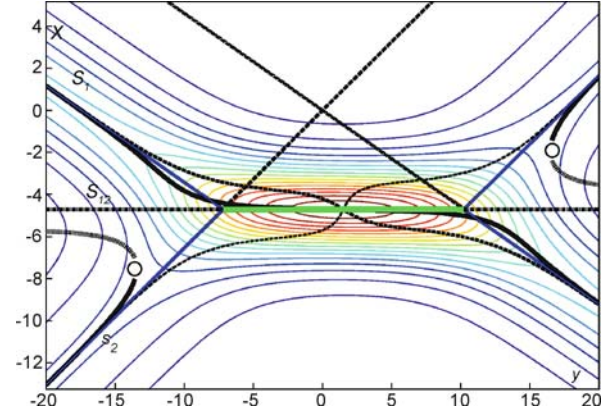
$$\begin{aligned} s_{1,2} &= \sqrt{A_{12}} k_{1,2}^2 \Theta^{-2} \cosh \left( \varphi_{2,1} x + \ln \sqrt{A_{12}} \right), \\ s_{12} &= 2\Theta^{-2} \left[ (k_1 - k_2)^2 + A_{12} (k_1 + k_2)^2 \right], \\ \Theta &= \cosh \frac{\varphi_1 - \varphi_2}{2} + \cosh \frac{\varphi_1 + \varphi_2 + \ln A_{12}}{2}. \end{aligned} \quad (7)$$

Here  $\varphi_i = k_i x + l_i y + \omega_i t$ ,  $a_i = \frac{1}{2} k_i^2$ ,  $i = 1, 2$ , are the phases and amplitudes of the interacting solitons, the ‘frequencies’  $\omega_i$  satisfy the dispersion relation  $k_i \omega_i + k_i^4 + 3l_i^2 = 0$  of the linearized KP equation,  $A_{12} = [\lambda^2 - (k_1 - k_2)^2] / [\lambda^2 - (k_1 + k_2)^2]$  is the phase shift parameter and  $\lambda = l_1 k_1^{-1} - l_2 k_2^{-1}$ .

The two-soliton solution only exists in a part of the parameter space  $R^{(4)}(k_1, k_2, l_1, l_2)$ . An interaction may result in either the positive or the negative phase shifts  $\delta_{1,2} = -\ln A_{12} / |k_{1,2}|$  of the counterparts (Fig. 6). The negative phase shift (cf. Fig. 3) can be attributed to an attractive interaction and the positive phase shift to a repulsive interaction.

The residue  $s_{12}$  (with an amplitude  $a_{12}$ ) is at times considered as a virtual 2D “interaction soliton” [118], which grows into the resonant soliton at the limit of exact resonance  $A_{12} \rightarrow \infty$  (cf. Fig. 7). This virtual structure can be heuristically interpreted as supported by both dispersive and diffractive effects, the latter being balanced by the oblique motion of the interacting solitons.

The patterns of both idealized (Fig. 3) and factual (Fig. 7) wave crests are symmetric with respect to a particular point called interaction center, and are stationary in a properly moving coordinate frame. If the amplitudes of the solitons are equal, the pattern is equivalent to the reflection of the incoming soliton from the wall following the motion of the interaction center. For unequal amplitude solitons the equivalence is not complete because mass and energy flux occur through such a wall.



Solitons Interactions, Figure 7

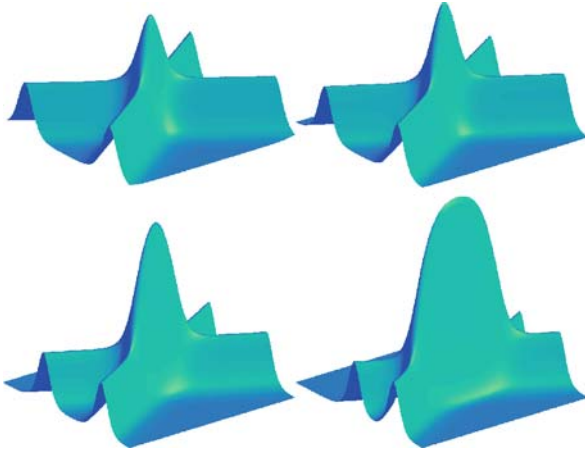
Patterns of idealized (solid/dashed straight lines) and factual (dashed curves) crests of the incoming solitons  $s_1, s_2$ , the crest of the residue  $s_{12}$  (green line), the visible wave crests and troughs (bold and bold dashed curves, respectively), and isolines of the two-soliton solution in the case  $l = l_1 = -l_2 = 0.3$ ,  $k_1 = 0.6507$ ,  $k_2 = 0.4599$  in normalized coordinates  $(x, y)$ . The interaction center is the crossing point of all factual crests. Contour interval is  $\frac{1}{4} \max(a_1, a_2) = 0.0689$ . Circles show the common points of the troughs and crests of the two-soliton solution. Reprinted from [146]

### Amplitudes

The phase shifts  $\delta_{1,2}$  of the counterparts only depend on the amplitudes of the incoming solitons and the angle  $\alpha_{12} = 2 \arctan(\frac{1}{2}\lambda)$  between their crests. For the negative phase shift case  $A_{12} > 1$  (that is typical in interactions of solitons with comparable amplitudes, Fig. 5), an interaction pattern emerges, the height of which exceeds the sum of the heights of two incoming solitons. The maximum surface elevation for equal amplitude solitons is  $a_{\max} = 4a_{1,2} / (1 + A_{12}^{-1/2})$ ; thus, the nonlinear superposition of solitons may lead to a fourfold amplification of the surface elevation. In a highly idealized case of interactions of five solitons the maximum surface elevation may exceed the amplitude of the incoming solitons by more than an order. The largest elevation occurs if the interacting solitons are in exact resonance  $A_{12} = \infty$ , when solitons intersect under a physical angle  $\tilde{\alpha}_{12} = 2 \arctan \sqrt{3\tilde{\eta}/\hbar}$ . For unequal amplitude solitons the maximum elevation  $a_{\max}$  for finite  $A_{12}$  and the amplitude of the resonant soliton  $a_{\infty}$  at  $A_{12} = \infty$  are [46,146]

$$a_{\max} = a_{12} + 2A_{12}^{1/2} \frac{a_1 + a_2}{(A_{12}^{1/2} + 1)^2}, \quad a_{\infty} = \frac{(k_1 + k_2)^2}{2}. \quad (8)$$

The near-resonant high hump is particularly narrow and its front is very steep. The maximum slope of the front of the two-soliton solution may be eight times as large as



**Solitons Interactions, Figure 8**

Surface elevation in the vicinity of the interaction area, corresponding to the incoming solitons with equal amplitudes intersecting under different angles [120]

the slope of the incoming solitons [150]. For unequal amplitude solitons, the amplification of slope of the front of the interaction pattern is roughly proportional to the amplitude amplification. The extraordinary steepness of the front of the near-resonant hump, although intriguing, is not unexpected, because the resonant soliton is higher and therefore narrower than the incoming solitons.

The length  $L_{12}$  of the idealized common crest (Fig. 3) is  $L_{12} \sim \ln A_{12}$  and therefore is modest unless the interacting solitons are near-resonant. The length of the area, where the total elevation exceeds the sum of the amplitudes of the counterparts, may be considerably longer; however it is also roughly proportional to  $\ln A_{12}$  [150]. For largely different amplitudes of the interacting solitons, the amplitude amplification remains modest but the spatial extent of substantial influence of nonlinear interaction is roughly as large as if the amplitudes were equal [151]. Their near-resonant interaction becomes evident as a wave with its crest oriented and propagating differently compared with the counterparts. The described features are universal for that part of the wave vector space, in which the two-soliton solution with a negative phase shift exists. In the case of a positive phase shift (that is typical for interactions of solitons with largely different amplitudes) the highest elevation does not exceed the amplitude of the larger soliton.

Numerical simulations of the process of formation of the composite structure show that a high wave hump is formed quite fast in the interaction region and its height soon considerably exceeds the sum of the heights of the counterparts [58,124]. A high and gradually lengthening wave hump is also formed in cases when no exact two-

soliton solution of the KP equations exists. The interaction in such cases is inelastic: neither the orientation of the crests nor the height of the solitons before and after interaction match each other, and a certain amount of wave radiation takes place. The described features become evident in a number of simulations of soliton interactions in different environments [105,168].

## Soliton Interactions in Laboratory and Nature

### Ion-Acoustic Solitons

The most productive studies of soliton interactions in the 1970s and the 1980s were performed in the framework of ion-acoustic waves. Their behavior is approximately described by the KP equation. An analytical solution describing interactions and resonance of two plane ion-acoustic solitons of small amplitude in the 3D collisionless plasma was given in [171]. This solution is wider than the two-soliton solution of the KP equation in the sense that the nearly unidirectional approximation is not used.

The first successful experiment proving that the 2D interaction of solitons has a wider spectrum of features compared with the 1D collisions is reported in [178]. A collision of two equal amplitude spherical solitons led to significant changes and to the emergence of a new 2D object. This interaction was not perfectly elastic: the solitons were not only shifted in phase but also reduced in amplitude. Soon afterwards, a near-resonant interaction of two KP solitons, with a large-amplitude wave hump of the counterparts oriented across the principal propagation direction, was observed near the interaction center, whereas the counterparts experienced clear phase shifts [43]. The amplitude amplification was found to be close to twofold the sum of the amplitudes of the incoming solitons [103].

The first adequate evidence of the nonlinear increase in amplitude during collisions of unequal amplitude KP solitons was also obtained in this framework [46]. For largely different amplitudes the measured maximum amplitudes were less than the predicted ones. A probable reason for the discrepancy is that the near-resonant wave hump occurs and covers a substantial area (equivalently, can be reliably estimated with the use of relatively large probes) only if the incoming solitons are very close to resonance. The common part of the crests of interacting solitons in the reported experiments was pretty short suggesting that the solitons were not exactly in resonance. Moreover, the complete identification of all the crests in the interactions of solitons with considerably different amplitudes is not possible because the crests are partially invisible (Fig. 7, [146]). The generic reason for experimental difficulties in this and similar experiments is the above-

discussed feature that large-amplitude structures tend to mask the presence of shorter ones.

### Shallow-Water and Internal Solitons

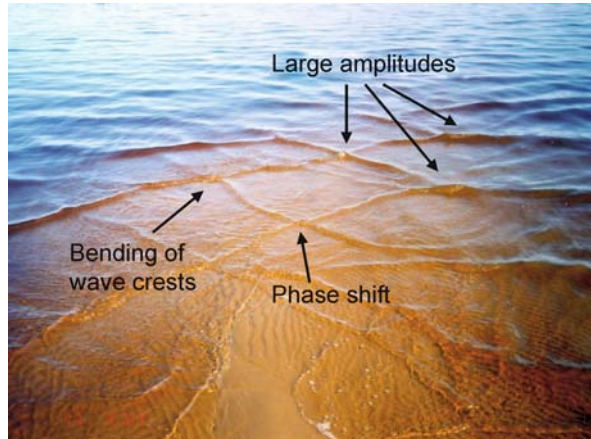
Phase shifts, formation of a common crest of the interacting soliton-like waves, and accompanying amplitude and orientation changes are commonly observable in very shallow areas (Fig. 9). A famous photo by Terry Toedtemeier (1978), in which two solitonic shallow-water waves have significant phase shifts and quite a long common crest on a beach in the US state of Oregon, has been reproduced in a number of sources (e. g. p. 24 in [34]).

Although [93] predicted up to fourfold amplification of the surface elevation, much weaker amplitude amplification was found in early experiments [91]. An amplification close to the theoretical one was detected as a by-product of studies of shallow-water channel superconductivity [25].

Russell's "great wave of translation" [132] probably had a practically straight crest, reflecting a well-known feature of the generation of solitons in relatively narrow channels: a nearly perfectly 1D upstream soliton is created also by a 2D disturbance. The KdV model, which is frequently used to describe this phenomenon, is intrinsically 1D and certainly results in straight wave crests. An analysis of the simplified problem of ship motion in the KP equation and the Boussinesq equation shows, in accordance with common experience, that the wave crests are curved [71]. In fact, the essentially 2D waves with paraboloidal crests generally emerge in front of the disturbance. Their curvature monotonously decreases as the waves propagate [85,87].

Although the sidewall of the channel is not essential for the radiation of the upstream solitons, still it plays a crucial role in the transformation of already radiated waves. The straight-crested solitons are formed in the process of the (Mach) reflection of solitons with curved crests from the sidewalls [113]. This feature can be distinguished in many numerical studies, where an initially curved soliton starts to straighten when it comes in touch with the sidewall [149]. Since the reflection of solitons, which in this case is equivalent to soliton interactions, plays the key important role in creating straight-crested waves in channels, with a certain exaggeration it can be said that Russell's soliton exists due to soliton interactions.

Probably the most common solitonic phenomenon in nature next to the shallow water waves is represented by internal solitary waves in the oceans and in the atmosphere ([56,107,111,130], see also the entry ► [Non-linear Internal Waves](#)). In many cases, the KdV equation is an appropriate model for internal waves and the effects occurring



**Solitons Interactions, Figure 9**

Interaction patterns of solitonic surface waves in very shallow water near Kauksi resort on Lake Peipsi, Estonia. Photo and copyright by Lauri Ilison, July 2003. First published in [151]

during their interactions are equivalent to those described above.

Large-scale internal solitary waves and their groups exhibit many features unique to soliton interactions and are often called simply internal solitons. Intersecting solitary wave packets on Synthetic Aperture Radar (SAR) images frequently exhibit phase shifts while crossing each other [51], similarly to phase shifts occurring in the KP model, where the agreement between theory and experiments is quite good. Apel et al. [9] summarize advances in the descriptions and observations of internal solitons and their interactions.

The other environment for internal solitons, which provides some instructive aspects and is eligible, e. g., for the description of small-scale internal solitons in a two-layer medium, is the Benjamin–Ono (BO) equation [1]

$$\eta_t + c\eta_x + \alpha\eta\eta_x + \frac{\beta}{\pi} \frac{\partial^2}{\partial x^2} \text{P.V.} \left[ \int_{-\infty}^{\infty} \frac{\eta(x', t)}{x - x'} dx' \right] = 0. \quad (9)$$

First, the BO solitons do not acquire a phase shift after a collision. Second, this environment demonstrates that generalization of integrable 1D soliton-admitting equations to even a weakly 2D case may lead to the loss of integrability, thus the preservation of this feature in the KP equation is not universal. The weakly 2D analogues of the BO equation [1] apparently are not integrable [9] and do not admit simple analytic multi-soliton solutions. Numerically simulated oblique interactions of weakly 2D BO solitons show that a phenomenon resembling the resonant



interaction of the KP solitons and an accompanied amplitude amplification do occur but the newly generated wave in the zone of nonlinear interaction is far from the BO soliton [105,167].

### Solitons at Planetary Scale

Large-scale internal waves and solitons in nature are affected by the joint influence of the Earth's rotation and sphericity, and perfect internal line solitons generally cannot exist in the ocean. The physical reason of this "antisoliton theorem" is that due to rotational dispersion, there is always a phase synchronism between a source moving at an arbitrary speed and linear perturbations in such environments. This leads to unavoidable wave radiation from a large-scale solitary internal wave [49]; however this effect is negligible for relatively small-scale solitons, the typical time scales of which in the oceans are a few tens of minutes.

The existence of nonstationary solitary waves is still possible [52]. They exhibit a phenomenon similar to the recurrence of solitons – the repeating decay and reemergence process, and formation of a nearly localized wave packet consisting of a long-wave envelope and shorter, faster solitary-like waves that propagate through the envelope [61]. The robust, long-lived structure may contain as much as 50% of the energy in the initial solitary wave. Interacting packets may either pass through one another, or merge to form a longer packet. Their further studies may reveal interesting features concerning the functioning of oceans and other large stratified water bodies. Yet in the majority of practically interesting cases such long-term behavior is masked by topographic effects that modify the internal waves much more dramatically.

Another medium of particular practical importance as well as an instructive one from the viewpoint of soliton interactions is the large-scale geophysical flow in the oceans and the atmosphere. It is frequently treated as anisotropic quasi-2D turbulence, which has the property of energy concentration in large-scale structures. Eddies in such flows are stabilized by the background rotation and also affected by the joint effect of the Earth's rotation and sphericity. The latter becomes evident through the North-South variation of the Coriolis force, the so-called (planetary) beta-effect. In this environment, large-scale vortices often maintain their coherence for surprisingly long times and show an amazing variety of interactions; perhaps the most well-known example being the Jovian Great Red Spot.

Form-preserving, uniformly translating, horizontally localized solutions (of the equations of planetary dy-

namics) are called modons. In nondissipative quasi-geostrophic dynamics they exhibit a variety of properties of solitons. They carry a certain amount of fluid with them, which distinguishes them from (solitary) waves that cause, if at all, a finite displacement of the medium. The word modon was coined a decade after soliton was first used [157].

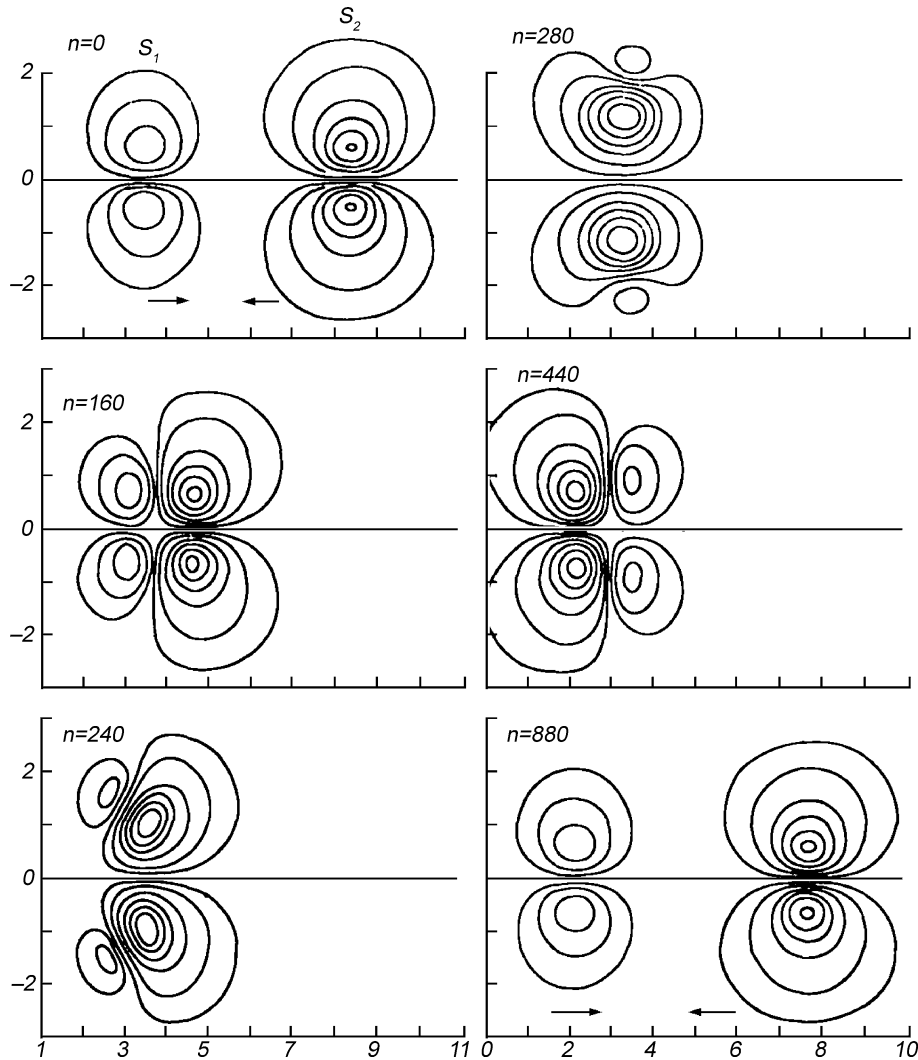
The standard quasi-geostrophic models predict that the accompanying Rossby-wave radiation should cause a decay of isolated eddies. Several models explain the persistence of monopolar vortices using external features of the flow, or interactions with a background shear that suppresses the Rossby-wave radiation [19,126,144]. A number of generalizations of the quasi-geostrophic model predict that, within a certain range of parameters, a nonlinear anticyclonic vortex of a special form can exist for a long time [121,153,170] due to the mutual compensation of weak (scalar) nonlinearity and weak dispersion. Some authors call such structures (Petviashvili-type) Rossby solitons, although their collisions are not elastic.

Long-living coherent vortices and their interactions have been studied in many experiments with rapidly rotating parabolic vessels [102]. The increase of the effective gravity due to the centrifugal acceleration towards the side walls modifies the governing equations, so that the (Petviashvili-type) Rossby soliton does not exist in the paraboloidal geometry [104]; yet anticyclonic vortices have a very long lifetime [154]. Another option to study the evolution of such vortices is a tank with a sloping bottom to imitate the planetary beta-effect (e.g. [41,90], among others; for an overview of earlier studies see [63]).

The majority of interactions of solitary vortices in such an environment are inelastic. The 2D localized vortices of comparable size and intensity, and like sign attract each other and generally coalesce soon. Only practically equivalent vortices may form a quasi-stable pair rotating around each other. The vortices of different signs either repulse from each other or form a relatively stable dipole [54,55]. The simultaneous evolution of a large number of soliton-like vortices has mostly been studied theoretically or numerically (e.g. in terms of a cluster of point vortices or hetons, [57]).

The simplest stable barotropic modon in flat-bottom beta-plane dynamics is the Larichev–Reznik dipole [83]. It is a traveling dipolar vortex with a characteristic north/south antisymmetry. It propagates eastward at any speed, or westward at speeds greater than the longest-wave speed. The speed of translation is therefore out of the range of the phase speed of Rossby waves, which is the reason why this modon does not radiate waves (unlike large-scale internal solitons). The significance of such structures in everyday





**Solitons Interactions, Figure 10**  
 Head-on collision of Larichev–Reznik dipoles [67]

life and climate dynamics (where they may be responsible, e. g., for certain unusual weather events since the vortex pair may behave completely differently compared with a single cyclone or anticyclone) is far from being well understood.

A part of the numerically simulated direct head-on and overtaking collisions of Larichev–Reznik modons are elastic [84]. An instructive feature of the head-on collisions is that the identity of all four vortices can be sometimes tracked during all the interaction phase (Fig. 10). The process resembles a collision of soft particle pairs connected by an elastic chain. Their oblique collisions generally are not elastic and may lead to the formation of new dipoles, tripoles, or to the destruction of the structures.

## Effects in Higher Dimensions

### Optical Spatial Solitons

Starting from the mid-1980s, various optical solitary entities have become a key arena for studies of the properties and interactions of solitary structures in multiple dimensions. The interest in these studies is continuously supported by a variety of existing and emerging applications of such solitons and their interactions in communication and computation technology (see also the entry ► [Optical Computing](#)). Another reason for the rapid progress in these studies is the relative ease of their generation in suitable materials, combined with the precision of contemporary optical experiments and the possibilities of precise

control over almost every parameter [136,152]. They offered a lot of the further conceptual progress of soliton interactions through making possible 3D propagation of solitary entities in laboratory conditions, albeit in many cases these entities did not match the classical nomenclature of solitons. They were called solitary waves in the field for several decades, because the governing equations are not necessarily integrable and the interactions not necessarily elastic. Still their interaction at times shows amazingly elastic nature and they are commonly called optical spatial solitons (OSS) in the modern nonlinear optics nomenclature [136].

A large number of various kinds of OSSs have been identified [76,135]. Observations of spatial Kerr solitons became possible from the mid-1980s, when so-called slab waveguides were built [3,13]. Their behavior substantially depends on the dimensions of the soliton and the waveguide. For example, the bright Kerr solitons are stable only in planar systems whereas the 2D solitons were found to collapse. Quasi-stable 2D solitons in Kerr media described by the cubic NLS in two spatial dimensions exist in the form of so-called necklace-ring beams [143].

Optical spatial solitons occurring due to the saturation of the nonlinear change in the refractive index were first demonstrated in the 1970s [18] and realized in a solid medium in the 1990s [74]. The net effect of the multiple physical effects involved in photorefractive materials is that the underlying nonlinearities are also saturable. Such solitons were also discovered in the early 1990s [137] and were found to be stable in both slab waveguides [75] and in bulk media [38].

So-called quadratic solitons (that consist of multifrequency waves coupled by means of a second-order nonlinearity) were predicted in the mid-1970s [70] and realized experimentally in the 1990s [164]. They can be thought of as vector solitons because they involve mutual self-trapping of two or more components [152]. Their interactions are similar to those occurring in other saturable nonlinear media.

### Coherent and Incoherent Collisions

Coherent interactions of OSSs occur when the nonlinear medium responds quickly to the (interference) effects in the overlapping area of the beams. The increase of the intensity of light in the overlapping region of two parallel launched equivalent in-phase solitons in the focusing medium leads to an increase in the refractive index in that region. This in turn forces the centroid of each soliton toward it. Hence the solitons attract each other. The spatial distribution of the intensity of light resembles the tem-

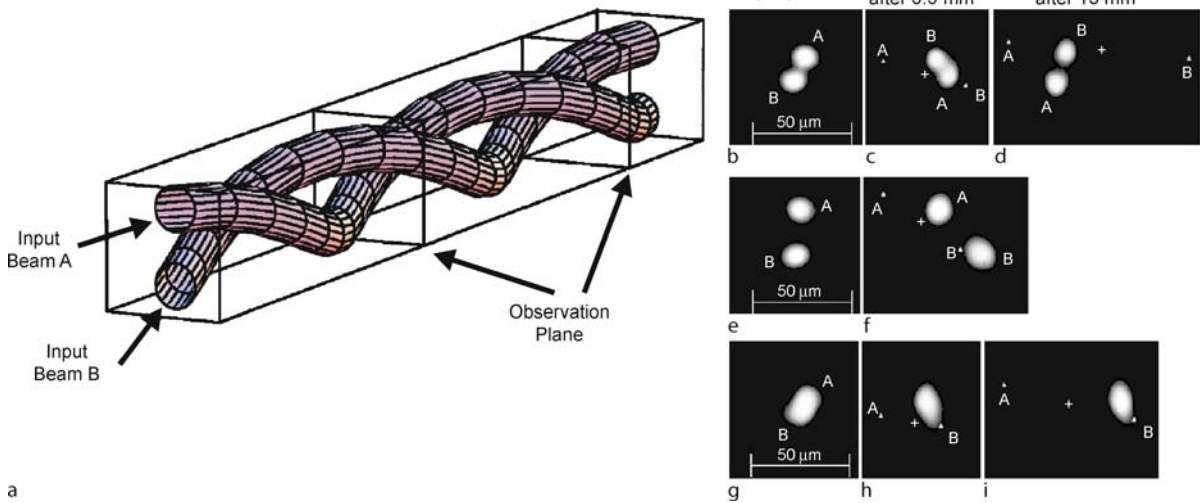
poral distribution of the amplitude of attracting interacting line solitons. In the case of antiphase beams, an opposite process occurs and such solitons exert repulsion. Coherent collisions of Kerr solitons have been demonstrated and the attraction for in-phase and the repulsion for antiphase solitons were clearly observed at the beginning of the 1990s [4,5,139]. The collision of nonequivalent coherent OSSs results in an energy exchange between the beams and in a repulsive force that makes the beams diverge.

Some reactions (e.g. photorefractive and thermal) of the optical medium are relatively slow. Solitons within which the phase varies randomly across the beam were first demonstrated in [95] and later extended to the beams of incoherent white light [94]. So-called incoherent soliton interactions occur when the response of the medium only follows the average intensity of light. Interactions of incoherent bright solitons are always attractive (Andersen and Lisak 1995).

The long-distance behavior of the counterparts in attractive interactions depends on the properties of the medium and the orientation of the beams. Pure Kerr solitons launched along nearly parallel directions result in periodic paths of the solitons' centroids. As in the case of the recurrence of the KdV solitons, Kerr solitons exactly return to their initial conditions after each cycle. Solitons launched under large enough converging angles exhibit a slight lateral deflection (which is equivalent to the phase shift of the KdV solitons). The solitons launched under large enough diverging angles never collide.

If the energy transfer were neglected, the force between the equivalent 1D Kerr solitons would vary from the maximum attractive between in-phase solitons to the maximum repulsive for antiphase solitons. This process leads to the (generally nonperiodic) energy transfer, where one soliton may gather net energy from the other. The details of the trajectories and other properties of the interacting solitons can be quite complex [152].

Interactions of 2D OSSs (e.g. in media with saturating nonlinearities) have a much larger variety of scenarios [152]; for example, interactions of incoherent photorefractive solitons may be both attractive and repulsive [156]. A number of resonance-like inelastic phenomena have been observed in which the number of solitons is not conserved. The *fusion* of two or more solitons into one structure resembles the resonance phenomenon. Here it happens owing to the gradual change of the orientation of the waveguides of attracting solitons launched under small relative angles into a new waveguide [48]. Differently from the resonance of line solitons, this happens for a certain range of initial parameters. The threshold is the maximum total internal reflection angle in the induced



**Solitons Interactions, Figure 11**

**a** The experimentally observed soliton spiraling process. The arrows indicate the initial trajectories. **b, e, and g** show different input conditions and **c, f, and h** are the outputs after 6.3 mm and **d and i** after 13 mm. The triangles indicate the centers of the corresponding diffracting beams. From [152]

waveguide [140,142]. For larger collision angles, the solitons interact as described above. The solitons can fuse together already on the first merging or after a certain number of oscillations of their trajectories with decreasing amplitudes and periods [12,78,160,161]. A sort of inverse of this process is the breakup (*fission*, [142]) of optical solitons into new solitons upon their interaction [163]. *Anihilation* of solitons upon collision also may occur, when three solitons collided and only two emerged from the collision process [79].

### Three-Dimensional Effects

In 3D space the trajectories of interacting beam solitons do not necessarily lie on a single plane and the system possesses nonzero initial angular momentum. The trajectories of solitons that are launched along skew lines passing close to each other may bend in three dimensions. The solitons *spiraling* around each other form a double helix orbit [123], much like two celestial objects or two moving charged particles moving along nearly parallel trajectories do. The interaction generally leads to spiraling-fusion or spiraling-repulsion of the trajectories [14,160,161]. Since the mutual rotation encounters a centrifugal force (which is always repulsive), a perfect double-helix DNA-like system, an interesting class of elastic interactions, requires soliton attraction. It may be formed from identical in-phase coherent solitons; yet such a double helix is

unstable with respect to perturbations of both the relative phase and amplitude of one of the solitons. This effect resembles processes occurring during the interactions of vortices of different and like sign in quasi-2D rotating fluids.

Attractive interactions of incoherent solitons have made possible the observation of spiraling solitons in saturable media [141]. The two solitons of equal power orbit periodically about each other. Such a pair conserves angular momentum and is more or less stable for a certain range of the initial orientation and distance between the beams. Although the underlying saturable nonlinearities are described by nonintegrable equations, the spiraling process does not lead to measurable energy radiation [20]. When the initial distance between the solitons is increased, they do not form a long-living interacting system. The double spiral for beams located initially very close to each other converges fast and the counterparts eventually fuse. The pair apparently has a limited lifetime. A likely reason for its fusion or off-spiraling is the potential coherence of the light in the beams.

### Interactions of Vector Solitons

Vector (composite) solitons consist of two or more components (also called modes) that mutually self-trap in a nonlinear medium. The simplest vector solitons, first suggested by Manakov [88], consist of two orthogonally

polarized components in a nonlinear Kerr medium in which self-phase modulation is identical to cross-phase modulation. They are described by an integrable system and interact elastically. An appropriate nonlinear material and experimental conditions for their demonstration were developed a long time after their discovery [68]. Similar solitons may exist if each field component is at a different frequency, and the frequency difference between components is much larger than the inverse of the nonlinearity relaxation time [138,166], or if the field components are incoherent with one another [22,30]. The energy exchange between the components of the interacting vector solitons takes place without radiative losses [7].

Vector solitons can also consist of different modes of their jointly induced waveguide [29,165]. These multimode solitons may possess multiple humps, may contain both bright and dark components [28,166], may have a certain internal dynamics, and were found to be stable (or weakly unstable) in large regions of their parameter space [96,109,110]. Interactions between vector solitons have, additionally to the above-discussed generic properties of solitons interactions (e.g. [86]), also some unique features. The shape transformations of colliding multimode solitons can lead to two different multimode solitons emerging from the collision process [6,69,80]. This “polarization switching” was predicted initially for Manakov solitons [125].

Composite multimode multihump solitons may carry topological charge in one of the vector components [21, 98]. This charge carried by a soliton of finite energy can be interpreted as spin of real particles. Elastic collisions of such 3D solitons should, ideally, conserve not only energy, and linear and angular momentum [122], but also the equivalent of spin. Two colliding composite solitons carrying opposite spins may form a metastable state that later decays into two or three new solitons. If the solitons interact under a certain critical angle, angular momentum is transferred from spin to orbital angular momentum. Finally, the shape transformation of the vortex component occurs at large collision angles, for which scalar solitons of all types simply go through one another unaffected or only have phase shifts [99,100]. The recent progress in optical vortices, vortex solitons, and their interactions is reviewed in [35].

### Applications of Line Soliton Interactions

A generic use of soliton interactions in every case where solitary waves or topological solitons may exist, follows from the definition of a soliton. It consists of establishing

if the structure is a solitonic one or not from the properties of their interactions.

The practical use of specific features of soliton interactions was started largely in parallel with the first technologies based on the use of solitons (e.g. optical soliton-based communications) more than a decade ago. The most promising is the technology of [Optical Computing](#) (see the relevant entry), which may be partially based on the interactions of vector solitons [155] and which may open completely new horizons in the architecture of computers.

### Soliton Interactions on a Water Surface

Several applications based on properties of elastic soliton interactions (or changes of certain fields during such interactions) have been proposed in the framework of water waves. From the purely geometrical properties of oblique interactions of plane solitons one may extract, e.g., the water wave height. The relations for the phase shifts and for the intersection angle can be reduced to the following transcendental equation:

$$\delta_1 \sqrt{2a_1 (1 + \lambda^2/4)} = \pm \ln \frac{\delta_2^2 \lambda^2 - 2(\delta_2 - \delta_1)^2 a_1^2}{\delta_2^2 \lambda^2 - 2(\delta_2 + \delta_1)^2 a_1^2}. \quad (10)$$

If the intersection angle  $\alpha_{12}$ , the phase shifts  $\delta_{1,2}$  and their signs (or the wave propagation directions) are known, equation (10) uniquely defines the heights of the interacting solitons [118,119].

The phenomenon of drastic increase of surface elevation and slope of wave fronts described in Sect. “[Geometry of Oblique Interactions of KP Line Solitons](#)” can be attributed to formation of “freak” or “giant” waves in the ocean, the height and steepness of which are considerably larger than expected based on the classical wave statistics [73]. Their sudden appearance is a feature of paramount importance to and a generic source of danger for navigation and in coastal and offshore engineering. Interactions of envelope (Schrödinger) solitons and breather solutions of the NLS [108] in deep water are their potential source, although the model based on envelope soliton interactions underestimates the probability of large wave formation compared with the fully nonlinear model [31]. Interacting solitary wave groups that emerge from a long wave packet can produce freak wave events and may lead to a threefold increase of the wave slope [32].

The problem of extremely large-amplitude internal waves is by no means less important, as they can pose acute (currently neither well understood nor accounted for) danger for submarines, oil and gas platforms, oil risers and pipelines, and to other engineering constructions

in relatively deep water. The problem of “freak” internal waves and their generation due to internal soliton interactions has only recently attracted the attention of researchers [81].

There is a clear potential in the use of the advances concerning the geometry of the line soliton interactions in studies of both internal and shallow water solitons. While phase shifts usually have no dynamical significance, unexpectedly large elevations, extreme slopes, or changes in orientation of wave crests owing to the (Mach) reflection or oblique interactions of solitonic waves frequently cause an acute danger. The particularly high (stem) waves may lead to hits by high waves arriving from an unexpected direction. They may break during propagation into shallow water and attack armor blocks from an unprotected side [89] or overtop the breakwaters in unexpected locations [172]. These effects may be particularly pronounced in the case of ship wakes that often approach seawalls or breakwaters from other directions than wind waves do [148]. The possibility of drastic steepening of the front of the near-resonant structure is immaterial in many physical systems but is a crucial component of danger from freak waves [73], since specifically steep and high waves present an acute danger (e.g. [162]). Even if the steepness of the wave front is not large in absolute terms, relatively steep and long, tsunami-like waves tend to become asymmetric and exhibit unexpectedly large runup heights [36].

The amplitude amplification, however, evidently becomes effective relatively seldom, because two or more systems of long-crested solitonic waves must approach a certain area from different directions, and extreme elevations and slopes only occur if the amplitudes of the interacting solitons, the angle between their crests and the water depth are specifically balanced. The long life-time of the resulting wave hump in favorable conditions [73] may drastically increase the probability of the occurrence of abnormally high waves.

### Ship-Induced Solitons and Wave Resistance

An increasing source of solitonic waves is the fast ship traffic in relatively shallow areas. The generation of shallow-water solitons is most effective for large ships sailing at speeds roughly equal to the maximum phase speed  $\sqrt{gh}$  of surface gravity waves [87]. The low decay rates and exceptional compactness of the solitonic ship wakes has led to a significant impact on the safety of people, property and craft, unusually high hydrodynamic loads in a part of the nearshore, and to a considerable remote impact of the ship traffic in shallow areas [112,147]. Such a wake has probably caused a fatal accident as far as about 10 km from the

sailing line already in 1912 [149]. The interactions of ship-induced solitons may lead to dangerous waves in the vicinity of coastal fairways and harbor entrances in otherwise sheltered areas [120].

A intriguing use of soliton interactions, combining effects occurring during an elastic oblique interaction of shallow-water solitons, followed by an annihilation of solitons and forced antisolitons, has been made for reducing the wave resistance in channels [25] and for catamaran design [26]. In a channel, the bow wave reflected from a side wall is used to cancel the stern wave, whereas the bow wave of one hull of the catamaran is used to cancel the stern wave of the other hull. Its practical use is possible for ships sailing at near-critical speeds, the wake of which may consist of a single soliton. The adequate calculation of the phase shift occurring during its Mach reflection (equivalently, during the interaction of two bow solitons) is a key component in achieving a perfect annihilation of properly timed bow solitons and the sterns' waves of depression.

The use of this effect by adjusting the channel width, the water depth and the ship's speed probably is the first intentional use of the features occurring during soliton interactions in the design of a certain technology [23,24]. At the exact design conditions the reflected bow wave completely cancels the stern wave and leads to a sort of channel superconductivity [25]. The same effect may occur between two ships moving in parallel [65].

Finally, it is interesting to note that already J.S. Russell may have been aware of this possibility. He describes efforts of a spirited horse which, pulling a boat in a canal, had drawn the boat up into its own wave leading to a significant reduction in resistance. The boat owner had noted this, and the event led to 'high-speed' service on some canals in the 1820s and 1830s [131]. We shall probably never know if the decrease of wave resistance occurred due to simple crossing of the critical speed or owing to the channel superconductivity; however, even the remote possibility that the practical use of soliton interactions may have been reported even before the first report about solitons, is remarkable.

### Future Directions

Although the theory of soliton interactions has been extensively developed during four decades and its basic features for low-dimensional environments have been well understood, it is still in the stage of rapid development. Its first practical applications appeared only a decade ago and there is evidently room for relevant developments in (geo)physical applications. Analysis of long-term evolution of ensembles of solitons is still a challenge and reveals



ever new interesting features even in the simplest soliton-admitting systems such as the KdV equation. Although several formal procedures of building explicit multi-soliton solutions to many equations have been known since the 1970s, studies of properties of (optionally resonant) interactions of several line solitons only started at the turn of the millennium. Many fascinating features of elastic interactions of solitons of different kinds in more complex systems will obviously become evident in the nearest future. The results in this direction may largely widen understanding of the integrability of the underlying equations.

Extremely rapid progress may be expected in studies into properties of long-lived soliton-like structures, in particular, of optical spatial solitons. The classification of their nomenclature and interactions is far from being completed. This environment is the key framework for identifying the new features of soliton interactions in higher dimensions, even though interactions of such structures are not necessarily elastic.

## Bibliography

1. Ablowitz MJ, Segur H (1981) Solitons and the inverse scattering transform. SIAM, Philadelphia
2. Ablowitz MJ, Kaup DJ, Newell AC, Segur H (1974) The inverse scattering transform – Fourier analysis for nonlinear problems. *Stud Appl Math* 53:249–315
3. Aitchison JS, Weiner AM, Silberberg Y, Oliver MK, Jackel JL, Leaird DE, Vogel EM, Smith PWE (1990) Observation of spatial optical solitons in a nonlinear glass wave-guide. *Opt Lett* 15(9):471–473
4. Aitchison JS, Silberberg Y, Weiner AM, Leaird DE, Oliver MK, Jackel JL, Vogel EM, Smith PWE (1991) Spatial optical solitons in planar glass wave-guides. *J Opt Soc Am B-Optical Phys* 8(6):1290–1297
5. Aitchison JS, Weiner AM, Silberberg Y, Leaird DE, Oliver MK, Jackel JL, Smith PWE (1991) Experimental-observation of spatial soliton-interactions. *Opt Lett* 16(1):15–17
6. Akhmediev N, Krolkowski W, Snyder AW (1998) Partially coherent solitons of variable shape. *Phys Rev Lett* 81(21):4632–4635
7. Anastassiou C, Segev M, Steiglitz K, Giordmaine JA, Mitchell M, Shih MF, Lan S, Martin J (1999) Energy-exchange interactions between colliding vector solitons. *Phys Rev Lett* 83(12):2332–2335
8. Anderson D, Lisak M (1985) Bandwidth limits due to incoherent soliton interaction in optical-fiber communication-systems. *Phys Rev A* 32(4):2270–2274
9. Apel JR, Ostrovsky LA, Stepanyants YA, Lynch JF (2007) Internal solitons in the ocean. *J Acoust Soc Am* 121(2):695–722
10. Arnold JM (1998) Varieties of solitons and solitary waves. *Opt Quantum Electron* 30:631–647
11. Askar'yan GA (1962) Effects of the gradient of strong electromagnetic beam on electrons and atoms. *Sov Phys JETP* 15(6):1088–1090
12. Baek Y, Schiek R, Stegeman GI, Baumann I, Sohler W (1997) Interactions between one-dimensional quadratic solitons. *Opt Lett* 22(20):1550–1552
13. Barthelemy A, Maneuf S, Froehly C (1985) Soliton propagation and self-confinement of laser-beams by Kerr optical nonlinearity. *Opt Commun* 55(3):201–206
14. Belic MR, Stepken A, Kaiser F (1999) Spiraling behavior of photorefractive screening solitons. *Phys Rev Lett* 82(3):544–547
15. Berger V, Kohlhasse S (1976) Mach-reflection as a diffraction problem. In: *Proc 25th Int Conf on Coastal Eng. ASCE, New York*, pp 796–814
16. Biondini G, Chakravarty S (2006) Soliton solutions of the Kadomtsev–Petviashvili II equation. *J Math Phys* 47(3):033514
17. Biondini G, Kodama Y (2003) On a family of solutions of the Kadomtsev–Petviashvili equation which also satisfy the Toda lattice hierarchy. *J Phys A-Math General* 36(42):10519–10536
18. Bjorkholm JE, Ashkin A (1974) cw self-focusing and self-trapping of light in sodium vapor. *Phys Rev Lett* 32(4):129–132
19. Busse FH (1994) Convection driven zonal flows and vortices in the major planets. *Chaos* 4:123–134
20. Buryak AV, Kivshar YS, Shih MF, Segev M (1999) Induced coherence and stable soliton spiraling. *Phys Rev Lett* 82(1):81–84
21. Carmon T, Anastassiou C, Lan S, Kip D, Musslimani ZH, Segev M, Christodoulides D (2000) Observation of two-dimensional multimode solitons. *Opt Lett* 25(15):1113–1115
22. Chen ZG, Segev M, Coskun TH, Christodoulides DN (1996) Observation of incoherently coupled photorefractive spatial soliton pairs. *Opt Lett* 21(18):1436–1438
23. Chen X-N, Sharma SD (1994) Nonlinear theory of asymmetric motion of a slender ship in a shallow channel. In: Rood EP (ed) 20th Symposium on Naval Hydrodynamics. US Office on Naval Research, Santa Barbara, pp 386–407
24. Chen X-N, Sharma SD (1997) Zero wave resistance for ships moving in shallow channels at supercritical speeds. *J Fluid Mech* 335:305–321
25. Chen X-N, Sharma SD, Stuntz N (2003) Zero wave resistance for ships moving in shallow channels at supercritical speeds. Part 2. Improved theory and model experiment. *J Fluid Mech* 478:111–124
26. Chen X-N, Sharma SD, Stuntz N (2003) Wave reduction by S-catamaran at supercritical speeds. *J Ship Res* 47:145–154
27. Chiao RY, Garmire E, Townes CH (1964) Self-trapping of optical beams. *Phys Rev Lett* 13(15):479–482
28. Christodoulides DN (1988) Black and white vector solitons in weakly birefringent optical fibers. *Phys Lett A* 132(8–9):451–452
29. Christodoulides DN, Joseph RI (1988) Vector solitons in birefringent nonlinear dispersive media. *Opt Lett* 13(1):53–55
30. Christodoulides DN, Singh SR, Carvalho MI, Segev M (1996) Incoherently coupled soliton pairs in biased photorefractive crystals. *Appl Phys Lett* 68(13):1763–1765
31. Clamond D, Grue J (2002) Interaction between envelope solitons as a model for freak wave formations. Part I: Long time interaction. *Comptes Rendus Mecanique* 330(8):575–580
32. Clamond D, Francius M, Grue J, Kharif C (2006) Long time interaction of envelope solitons and freak wave formations. *Eur J Mech B-Fluids* 25(5):536–553
33. Davis RE, Acrivos A (1967) Solitary internal waves in deep water. *J Fluid Mech* 29(3):593–608

34. Dauxois T, Peyrard M (2006) *Physics of solitons*. Cambridge University Press, Cambridge
35. Desyatnikov AS, Kivshar YS, Torner L (2005) Optical vortices and vortex solitons. *Prog Opt* 47:291–391
36. Didenkulova II, Zahibo N, Kurkin AA, Levin BV, Pelinovsky EN, Soomere T (2006) Runup of nonlinearly deformed waves on a coast. *Doklady Earth Sci* 411(8):1241–1243
37. Drazin PG, Johnson RS (1989) *Solitons: An introduction*. In: Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge
38. Duree GC, Shultz JL, Salamo GJ, Segev M, Yariv A, Crosignani B, Diporto P, Sharp EJ, Neurgaonkar RR (1993) Observation of self-trapping of an optical beam due to the photorefractive effect. *Phys Rev Lett* 71(4):533–536
39. Engelbrecht J, Salupere A (2005) On the problem of periodicity and hidden solitons for the KdV model. *Chaos* 15:015114
40. Fermi E, Pasta J, Ulam S (1955) Studies of nonlinear problems. I Los Alamos report LA-1940; (1965) In: Segré E (ed) *Collected papers of Enrico Fermi*. University of Chicago Press, Chicago
41. Firing E, Beardsley RC (1976) The behaviour of a barotropic eddy on a beta-plane. *J Phys Oceanogr* 6:57–65
42. Flierl GR, Larichev VD, McWilliams JC, Reznik GM (1980) The dynamics of baroclinic and barotropic solitary eddies. *Dyn Atmos Oceans* 5(1):1–41
43. Folkes PA, Ikezi H, Davis R (1980) Two-dimensional interaction of ion-acoustic solitons. *Phys Rev Lett* 45(11):902–904
44. Fuerst RA, Canva MTG, Baboiu D, Stegeman GI (1997) Properties of type II quadratic solitons excited by imbalanced fundamental waves. *Opt Lett* 22(23):1748–1750
45. Funakoshi M (1980) Reflection of obliquely incident large-amplitude solitary wave. *J Phys Soc Japan* 49:2371–2379
46. Gabl EF, Lonngren KE (1984) On the oblique collision of unequal amplitude ion-acoustic solitons in a field-free plasma. *Phys Lett A* 100:153–155
47. Gardner CS, Greene JM, Kruskal MD, Miura RM (1974) Korteweg–de Vries equations and generalizations: methods for exact solutions. *Commun Pure Appl Math* 27:97–133
48. Gatz S, Herrmann J (1992) Soliton collision and soliton fusion in dispersive materials with a linear and quadratic intensity depending refraction index change. *IEEE J Quantum Electron* 28(7):1732–1738
49. Galkin VM, Stepanyants YA (1991) On the existence of stationary solitary waves in a rotating fluid. *PMM J Appl Math Mech* 55(6):939–943
50. Gardner CS, Green JM, Kruskal MD, Miura RM (1967) Method for solving the Korteweg–de Vries equation. *Phys Rev Lett* 19:1095–1097
51. Gasparovic RF, Apel JR, Kasischke ES (1988) An overview of the SAR internal wave signature experiment. *J Geophys Res-Oceans* 93(C10):12304–12316
52. Gilman OA, Grimshaw R, Stepanyants YA (1996) Dynamics of internal solitary waves in a rotating fluid. *Dyn Atmos Oceans* 23(1–4):403–411
53. Gordon JP (1983) Interaction forces among solitons in optical fibers. *Opt Lett* 8(11):596–598
54. Griffiths RW, Hopfinger EJ (1986) Experiments with baroclinic vortex pairs in a rotating fluid. *J Fluid Mech* 173:501–518
55. Griffiths RW, Hopfinger EJ (1987) Coalescing of geostrophic vortices. *J Fluid Mech* 178:73–97
56. Grimshaw R (2001) Internal solitary waves. In: Grimshaw R (ed) *Environmental Stratified Flows*. Kluwer, Dordrecht, pp 1–30
57. Gryanik VM, Doronina TN, Olbers DJ, Warncke TH (2000) The theory of three-dimensional hetons and vortex-dominated spreading in localized turbulent convection in a fast rotating stratified fluid. *J Fluid Mech* 423:71–125
58. Haragus-Courcelle M, Pego RL (2000) Spatial wave dynamics of steady oblique wave interactions. *Physica D* 145:207–232
59. Hasegawa A (1989) *Optical Solitons in Fibers*. Springer, Berlin
60. Hasegawa A, Tappert F (1973) Transmission of stationary nonlinear optical pulses in dispersive dielectric fiber: II Normal dispersion. *Appl Phys Lett* 23:171–172
61. Helfrich KR (2007) Decay and return of internal solitary waves with rotation. *Phys Fluid* 19(2):026601
62. Hirota R (1971) Exact solution of the Korteweg–de Vries equation for multiple collisions of solitons. *Phys Rev Lett* 27:1192–1194
63. Hopfinger EJ, van Heijst GJF (1993) Vortices in rotating fluids. *Annu Rev Fluid Mech* 25:241–289
64. Ikezi H, Taylor RJ, Baker DR (1970) Formation and interaction of ion-acoustic solitons. *Phys Rev Lett* 25(1):11–14
65. Jiankang W, Lee TS, Chu C (2001) Numerical study of wave interaction generated by two ships moving parallelly in shallow water. *Comput Meth Appl Mech Engrg* 190:2099–2110
66. Kadomtsev BB, Petviashvili VI (1970) The stability of solitary waves in weakly dispersive media. *Dokl Akad Nauk SSSR* 192:532–541
67. Kamenkovich et al. (1987) Synoptic eddies in the ocean. Russian version. *Gidrometeoizdat, Leningrad*, pp 124
68. Kang JU, Stegeman GI, Aitchison JS, Akhmediev N (1996) Observation of Manako spatial solitons in AlGaAs planar waveguides. *Phys Rev Lett* 76(20):3699–3702
69. Kanna T, Lakshmanan M (2003) Exact soliton solutions of coupled nonlinear Schrödinger equations: Shape-changing collisions, logic gates, and partially coherent solitons. *Phys Rev E* 67(4):046617 Part 2
70. Karamzin YN, Sukhorukov AP (1976) Mutual focusing of high-power light. *Sov Phys JETP* 41:414–420
71. Katsis C, Akylas TR (1987) On the excitation of long nonlinear water waves by a moving pressure distribution. Part 2. Three-dimensional effects. *J Fluid Mech* 177:49–65
72. Kaup DJ, Newell AC (1978) Solitons as particles, oscillators, and in slowly changing media: A singular perturbation theory. *Proc R Soc London A* 361(4):413–446
73. Kharif C, Pelinovsky N (2003) Physical mechanisms of the rogue wave phenomenon. *Eur J Mech B Fluids* 22:603–634
74. Khitrova G, Gibbs HM, Kawamura Y, Iwamura H, Ikegami T, Sipe JE, Ming L (1993) Spatial solitons in a self-focusing semiconductor gain medium. *Phys Rev Lett* 70(7):920–923
75. Kip D, Wesner M, Shandarov V, Moretti P (1998) Observation of bright spatial photorefractive solitons in a planar strontium barium niobate waveguide. *Opt Lett* 23(12):921–923
76. Kivshar YS, Luther-Davies B (1998) Dark optical solitons: physics and applications. *Phys Rep – Rev Sect Phys Lett* 298(2–3):81–197
77. Kivshar YS, Malomed BA (1989) Dynamics of solitons in nearly integrable systems. *Rev Mod Phys* 61(4):768–915
78. Krolkowski W, Holmstrom SA (1997) Fusion and birth of spatial solitons upon collision. *Opt Lett* 22(6):369–377

79. Krolkowski W, Luther-Davies B, Denz C, Tschudi T (1998) Anihilation of photorefractive solitons. *Opt Lett* 23(2):97–99
80. Krolkowski W, Akhmediev N, Luther-Davies B (1999) Collision-induced shape transformations of partially coherent solitons. *Phys Rev E* 59(4):4654–4658
81. Kurkin AA, Pelinovsky EN (2004) Freak waves: Facts, theory and modelling. Nizhny Novgorod State Technical University, Nizhny Novgorod (in Russian)
82. Lakshmanan M, Rajasekhar S (2003) Nonlinear dynamics: integrability, chaos and patterns. Springer, Berlin
83. Larichev VD, Reznik GM (1976) Strongly non-linear, two-dimensional isolated Rossby waves. *Oceanology* 16:961–967
84. Larichev VD, Reznik GM (1983) Collision of two-dimensional solitary Rossby waves. *Oceanology* 23(5):725–734
85. Lee SJ, Grimshaw RHJ (1990) Upstream-advancing waves generated by three-dimensional moving disturbances. *Phys Fluids A* 2:194–201
86. Leo G, Assanto G (1997) Collisional interactions of vectorial spatial solitary waves in type II frequency-doubling media. *J Opt Soc Am B-Opt Phys* 14(11):3151–3161
87. Li Y, Sclavounos PD (2002) Three-dimensional nonlinear solitary waves in shallow water generated by an advancing disturbance. *J Fluid Mech* 470:383–410
88. Manakov SV (1974) On the theory of two-dimensional stationary self-focusing of electromagnetic waves. *Sov Phys JETP* 38:248–253
89. Mase H, Memita T, Yuhi M, Kitano T (2002) Stem waves along vertical wall due to random wave incidence. *Coast Eng* 44:339–350
90. Masuda A, Marubayashi K, Ishibashi M (1990) A laboratory experiment and numerical simulation of an isolated barotropic eddy in a basin with topographic  $\beta$ . *J Fluid Mech* 213:641–659
91. Melville WK (1980) On the Mach reflection of solitary waves. *J Fluid Mech* 98:285–297
92. Miles JW (1977) Obliquely interacting solitary waves. *J Fluid Mech* 79:157–169
93. Miles JW (1977) Resonantly interacting solitary waves. *J Fluid Mech* 79:171–179
94. Mitchell M, Segev M (1997) Self-trapping of incoherent white light. *Nature* 387(6636):880–883
95. Mitchell M, Chen ZG, Shih MF, Segev M (1996) Self-trapping of partially spatially incoherent light. *Phys Rev Lett* 77(3):490–493
96. Mitchell M, Segev M, Christodoulides DN (1998) Observation of multihump multimode solitons. *Phys Rev Lett* 80(21):4657–4660
97. Mollenauer LF, Stolen RH, Gordon JP (1980) Experimental-observation of picosecond pulse narrowing and solitons in optical fibers. *Phys Rev Lett* 45(13):1095–1098
98. Musslimani ZH, Segev M, Christodoulides DN, Soljacic M (2000) Composite multihump vector solitons carrying topological charge. *Phys Rev Lett* 84(6):1164–1167
99. Musslimani ZH, Soljacic M, Segev M, Christodoulides DN (2001) Interactions between two-dimensional composite vector solitons carrying topological charges. *Phys Rev E* 63(6):066608
100. Musslimani ZH, Soljacic M, Segev M, Christodoulides DN (2001) Delayed-action interaction and spin-orbit coupling between solitons. *Phys Rev Lett* 86(5):799–802
101. Newell AC, Redekopp LG (1977) Breakdown of Zakharov–Shabat theory and soliton creation. *Phys Rev Lett* 38(8):377–380
102. Nezlin VM, Sneshkin EN (1993) Rossby vortices, spiral structures, solitons: Astrophysics and plasma physics in shallow water experiments. Springer, Berlin
103. Nishida Y, Nagasawa T (1980) Oblique collision of plane ion-acoustic solitons. *Phys Rev Lett* 45(20):1626–1629
104. Nycander J (1993) The difference between monopole vortices in planetary flows and laboratory experiments. *J Fluid Mech* 254:561–577
105. Oikawa M, Tsuji H (2006) Oblique interactions of weakly nonlinear long waves in dispersive systems. *Fluid Dyn Res* 38:868–898
106. Osborne AR, Bergamasco L (1986) The solitons of Zabusky and Kruskal revisited: perspective in terms of the periodic spectral transform. *Physica D* 18:26–46
107. Osborne AR, Burch TL (1980) Internal solitons in the Andaman Sea. *Science* 208(4443):451–460
108. Osborne AR, Onorato M, Serio M (2000) The nonlinear dynamics of rogue waves and holes in deep-water gravity wave trains. *Phys Lett A* 275(5–6):386–393
109. Ostrovskaya EA, Kivshar YS, Chen ZG, Segev M (1999) Interaction between vector solitons and solitonic gluons. *Opt Lett* 24(5):327–329
110. Ostrovskaya EA, Kivshar YS, Skryabin DV, Firth WJ (1999) Stability of multihump optical solitons. *Phys Rev Lett* 83(2):296–299
111. Ostrovsky LA, Stepanyants YA (1989) Do internal solitons exist in the ocean? *Rev Geophys* 27:293–310
112. Parnell KE, Kofoed-Hansen H (2001) Wakes from large high-speed ferries in confined coastal waters: Management approaches with examples from New Zealand and Denmark. *Coastal Manage* 29:217–237
113. Pedersen G (1988) Three-dimensional wave patterns generated by moving disturbances at transcritical speeds. *J Fluid Mech* 196:39–63
114. Pedersen NF, Samuelsen MR, Welner D (1984) Soliton annihilation in the perturbed sine-Gordon system. *Phys Rev B* 30(7):4057–4059
115. Peregrine DH (1983) Wave jumps and caustics in the propagation of finite-amplitude water waves. *J Fluid Mech* 136:435–452
116. Perring JK, Skyrme THR (1962) A Model Unified Field Equation. *Nucl Phys* 31:550–555
117. Perroud PH (1957) The Solitary Wave Reflection Along a Straight Vertical Wall at Oblique Incidence. Univ. of Calif. Berkeley, IER Rept 99-3, pp 93
118. Peterson P, van Groesen E (2000) A direct and inverse problem for wave crests modelled by interactions of two solitons. *Physica D* 141:316–332
119. Peterson P, van Groesen E (2001) Sensitivity of the inverse wave crest problem. *Wave Motion* 34:391–399
120. Peterson P, Soomere T, Engelbrecht J, van Groesen E (2003) Soliton interaction as a possible model for extreme waves in shallow water. *Nonlinear Process Geophys* 10:503–510
121. Petviashvili VI (1980) Red Spot of Jupiter and the drift soliton in plasma. *JETP Lett* 32:619–622
122. Pigier C, Uzdin R, Carmon T, Segev M, Nepomnyashchy A, Musslimani ZH (2001) Collisions between (2 + 1)D rotating propeller solitons. *Opt Lett* 26(20):1577–1579

123. Poladian L, Snyder AW, Mitchell DJ (1991) Spiraling spatial solitons. *Opt Commun* 85(1):59–62
124. Porubov AV, Tsuji H, Lavrenov IV, Oikawa M (2005) Formation of the rogue wave due to non-linear two-dimensional waves interaction. *Wave Motion* 42:202–210
125. Radhakrishnan R, Lakshmanan M, Hietarinta J (1997) Inelastic collision and switching of coupled bright solitons in optical fibers. *Phys Rev E* 56(2):2213–2216
126. Read PL, Hide R (1983) Long-lived eddies in the laboratory and in the atmosphere of Jupiter and Saturn. *Nature* 302(10):126–129
127. Rebba C (1979) Solitons. *Sci Am* 240(2):76–91
128. Redekopp LG, Weidman PD (1978) Solitary Rossby waves in zonal shear flows and their interactions. *J Atm Sci* 35:790–804
129. Rotschild C, Alfassi B, Cohen O, Segev M (2006) Long-range interactions between optical solitons. *Nat Phys* 2(11):769–774
130. Rottman JW, Grimshaw R (2001) Atmospheric internal solitary waves. In: Grimshaw R (ed), *Environmental Stratified Flows*. Kluwer, Dordrecht, pp 91–129
131. Russell JS (1837) Applications and illustrations of the law of wave in the practical navigation of canals. *Trans R Soc Edin* 14:33–34
132. Russell JS (1844) Report on waves. In: Report of the 14th Meeting of the British Association for the Advancement of Science. Murray, York, pp 311–390
133. Salupere A, Peterson P, Engelbrecht J (2002) Long-time behaviour of soliton ensembles. Part I – Emergence of ensembles. *Chaos Solitons Fractals* 14(9):1413–1424
134. Salupere A, Peterson P, Engelbrecht J (2003) Long-time behaviour of soliton ensembles. Part II – Periodical patterns of trajectories. *Chaos Solitons Fractals* 15:29–40
135. Segev M (1998) Optical spatial solitons. *Opt Quantum Electron* 30(7–10):03–533
136. Segev M, Stegeman G (1998) Self-trapping of optical beams: Spatial solitons. *Phys Today* 51(8):42–48
137. Segev M, Crosignani B, Yariv A, Fischer B (1992) Spatial solitons in photorefractive media. *Phys Rev Lett* 68(7):923–926
138. Shalaby M, Barthelemy AJ (1992) Observation of the self-guided propagation of a dark and bright spatial soliton pair in a focusing nonlinear medium. *IEEE J Quantum Electron* 28(12):2736–2741
139. Shalaby M, Reynaud F, Barthelemy A (1992) Experimental-observation of spatial soliton-interactions with a  $\pi$ -2 relative phase difference. *Opt Lett* 17(11):778–780
140. Shih MF, Segev M (1996) Incoherent collisions between two-dimensional bright steady-state photorefractive spatial screening solitons. *Opt Lett* 21(19):1538–1540
141. Shih MF, Segev M, Salamo G (1997) Three-dimensional spiraling of interacting spatial solitons. *Phys Rev Lett* 78(13):2551–2554
142. Snyder AW, Sheppard AP (1993) Collisions, steering, and guidance with spatial solitons. *Opt Lett* 18(7):482–484
143. Soljacic M, Sears S, Segev M (1998) Self-trapping of “necklace” beams in self-focusing Kerr media. *Phys Rev Lett* 81(22):4851–4854
144. Sommeria J, Nore C, Dumont T, Robert R (1991) Statistical theory of the Great Red Spot of Jupiter. *C R Acad Sci II* 312:999–1005
145. Soomere T (1992) Geometry of the double resonance of Rossby waves. *Ann Geophys* 10:741–748
146. Soomere T (2004) Interaction of Kadomtsev–Petviashvili solitons with unequal amplitudes. *Phys Lett A* 332:74–81
147. Soomere T (2005) Fast ferry traffic as a qualitatively new forcing factor of environmental processes in non-tidal sea areas: a case study in Tallinn Bay, Baltic Sea. *Environ Fluid Mech* 5:4 293–323
148. Soomere T (2006) Nonlinear ship wake waves as a model of rogue waves and a source of danger to the coastal environment: a review. *Oceanologia* 48(5):185–202
149. Soomere T (2007) Nonlinear components of ship wake waves. *Appl Mech Rev* 60(3):120–138
150. Soomere T, Engelbrecht J (2005) Extreme elevations and slopes of interacting solitons in shallow water. *Wave Motion* 41:179–192
151. Soomere T, Engelbrecht J (2006) Weakly two-dimensional interaction of solitons in shallow water. *Eur J Mech B Fluids* 25(5):636–648
152. Stegeman GI, Segev M (1999) Optical spatial solitons and their interactions: Universality and diversity. *Science* 286(5444):1518–1523. doi:10.1126/science.286.5444.1518
153. Stegner A, Zeitlin V (1996) Asymptotic expansions and monopolar solitary Rossby vortices in barotropic and two-layer model. *Geophys Astrophys Fluid Dyn* 83:159–195
154. Stegner A, Zeitlin V (1998) From quasi-geostrophic to strongly non-linear monopolar vortices in a paraboloidal shallow-water experiment. *J Fluid Mech* 356:1–24
155. Steiglitz K (2001) Time-gated Manakov spatial solitons are computationally. *Phys Rev E* 63(1):016608
156. Stepken A, Kaiser F, Belic MR, Krolikowski W (1998) Interaction of incoherent two-dimensional photorefractive solitons. *Phys Rev E* 58(4):R4112–R4115
157. Stern ME (1975) Minimal properties of planetary eddies. *J Mar Res* 33:1–13
158. Tajiri M, Maesono H (1997) Resonant interactions of drift vortex solitons in a convective motion of a plasma. *Phys Rev E* 55(3):3351–3357
159. Tanaka M (1993) Mach reflection of a large-amplitude solitary wave. *J Fluid Mech* 248:637–661
160. Tikhonenko V, Christou J, Luther-Davies B (1995) Spiraling bright spatial solitons formed by the breakup of an optical vortex in a saturable self-focusing medium. *J Opt Soc Am B-Opt Phys* 12(11):2046–2052
161. Tikhonenko V, Christou J, Luther-Davies B (1996) Three dimensional bright spatial soliton collision and fusion in a saturable nonlinear medium. *Phys Rev Lett* 76(15):2698–2701
162. Toffoli A, Lefevre JM, Bitner-Gregersen E, Monbaliu J (2005) Towards the identification of warning criteria: Analysis of a ship accident database. *Appl Ocean Res* 27(6):281–291
163. Torner L, Torres JP, Menyuk CR (1996) Fission and self-deflection of spatial solitons by cascading. *Opt Lett* 21(7):462–464
164. Torruellas WE, Wang Z, Hagan DJ, Vanstryland EW, Stegeman GI, Torner L, Menyuk CR (1995) Observation of 2-dimensional spatial solitary waves in a quadratic medium. *Phys Rev Lett* 74(25):5036–5039
165. Tratnik MV, Sipe JE (1988) Bound solitary waves in a birefringent optical fiber. *Phys Rev A* 38(4):2011–2017
166. Trillo S, Wabnitz S, Wright EM, Stegeman GI (1988) Optical solitary waves induced by cross-phase modulation. *Opt Lett* 13(10):871–873



167. Tsuji H, Oikawa M (2001) Oblique interaction of internal solitary waves in a two-layer fluid of infinite depth. *Fluid Dyn Res* 29:251–267
168. Tsuji H, Oikawa M (2004) Two-dimensional interaction of solitary waves in a modified Kadomtsev–Petviashvili equation. *J Phys Soc Japan* 73:3034–3043
169. Washimi H, Taniuti T (1966) Propagation of ion-acoustic solitary waves of small amplitude. *Phys Rev Lett* 17:996–998
170. Williams GP, Yamagata T (1984) Geostrophic regimes, intermediate solitary vortices and Jovian eddies. *J Atm Sci* 41:453–478
171. Yajima N, Oikawa M, Satsuma J (1978) Interaction of ion-acoustic solitons in three-dimensional space. *J Phys Soc Japan* 44:1711–1714
172. Yoon SB, Liu PLF (1989) Stem waves along breakwater. *J Waterw Port Coast Ocean Eng – ASCE* 115:635–648
173. Yue DK, Mei CC (1980) Forward diffraction of Stokes waves by a thin wedge. *J Fluid Mech* 99:33–52
174. Zabusky NJ, Kruskal MD (1965) Interaction on “solitons” in a collisionless plasma and the recurrence of initial states. *Phys Rev Lett* 15(6):240–243
175. Zakharov VE (1968) Stability of periodic waves of finite amplitude on the surface of a deep fluid. *Zh Prikl Mekh Tekh Fiz* 9:86–94; *J Appl Mech Tech Phys* 9:190–194
176. Zakharov VE, Shabat AB (1972) Exact theory of two-dimensional self-focusing and one-dimensional self-modulation of waves in nonlinear media. *Sov Phys JETP* 34:62–69
177. Zakharov VE, Manakov SV, Novikov SP, Pitaevsky LP (1980) *Theory of solitons*. Nauka, Moscow (in Russian) (English translation, 1984, Consultants Bureau, New York)
178. Ze F, Hershkowitz N, Chan C, Lonngren KE (1979) Inelastic collision of spherical ion-acoustic solitons. *Phys Rev Lett* 42(26):1747–1750

## Solitons, Introduction to

MOHAMED A. HELAL

Department of Mathematics, Faculty of Science,  
Cairo University, Giza, Egypt

The concept of solitons reflects one of the most important developments in science at the second half of the 20th century: the nonlinear description of the world. It is not easy to give a comprehensive and precise definition of a soliton. Frequently, a *soliton* is explained as a spatially localized wave in a medium that can interact strongly with other solitons but will afterwards regain its original form.

It is a nonlinear pulse-like wave that can exist in some nonlinear systems. The soliton wave can propagate without dispersing its energy over a large region of space; collision of two solitons leads to unchanged forms, solitons also exhibit particle-like properties.

The most remarkable property of solitons is that they do not disperse and thus conserve their form during propagation and collision. ► [Solitons Interactions](#).

The nonlinear science has been growing for approximately fifty years. However, numerous nonlinear processes had been previously identified, but the nonlinear mathematical tools were not developed. The available tools were linear, and nonlinearities were avoided or treated as perturbations of linear theories.

The first experimental observation of a solitary wave was made in August 1834 by a Scottish engineer named John Scott Russell (1808–1882). Scott Russell reported his observation; to the Fourteenth Meeting of the British Association for Advancement of Science, held in 1844; in a long report entitled: “Report on waves” as follows:

*“I was observing the motion of a boat which was rapidly drawn along a narrow channel by a pair of horses, when the boat suddenly stopped – not so the mass of water in the channel which it had put in motion; it accumulated round the prow of the vessel in a state of violent agitation, then suddenly leaving it behind, rolled forward with great velocity, assuming the form of a large solitary elevation, a rounded, smooth and well-defined heap of water, which continued its course along the channel apparently without change of form or diminution of speed. I followed it on horseback, and overtook it still rolling on at a rate of some eight or nine miles an hour, preserving its original figure some thirty feet long and a foot to a foot and a half in height. Its height gradually diminished, and after a chase of one or two miles I lost it in the windings of the channel. Such, in the month of August 1834, was my first chance interview with that singular and beautiful phenomenon which I have called the Wave of Translation”.*

This hump-shape localized wave that propagates along one space-direction with undeformed shape has spectacular stability properties. John Scott Russell carried out many experiments to get the properties of this wave. ► [Solitons: Historical and Physical Introduction](#).

A model equation describing the unidirectional propagation of long waves in water of relatively shallow depth with a localized solution (representing a single hump as discovered by Russell) was obtained by Korteweg and de Vries. This equation has become very famous and is now known as the Korteweg–de Vries equation or KdV equation. ► [Water Waves and the Korteweg–de Vries Equation](#).

There exists a certain class of nonlinear partial differential equations that leads to solitons. The Korteweg–de Vries equation, Kadomtsev–Petviashvili (KP) equation, Klein–Gordon (KG) equation, Sine–Gordon (SG) equa-



tion, nonlinear Schrodinger (NLS) equation, Korteweg–de Vries Burger's (KdVB) equation, etc. . . are some known equations that lie in this specific class of this NLPDE. This class of nonlinear equations has a special type of the traveling wave solutions which are either solitary waves or solitons. ► [Partial Differential Equations that Lead to Solitons](#).

The Korteweg–de Vries equation is a canonical equation in nonlinear wave physics demonstrating the existence of solitons and their elastic interactions. This KdV equation attracted many authors to publish an enormous number of publications. Different analytical methods for solving this well-known equation are presented in ► [Korteweg–de Vries Equation \(KdV\), Different Analytical Methods for Solving the](#).

One of the most famous analytical methods for solving the KdV equation is the *Inverse Scattering Transformation*. This method was first introduced by the well-known scientists Gardener, Green, Kruskal, and Miura. In 1967, Gardner et al. published a magnificent and original paper on the analytical solution to the initial value problem due to a disturbance of finite amplitude in an infinite domain. This systematic method is not easy, and needs several long steps to get the complete solution. ► [Inverse Scattering Transform and the Theory of Solitons](#).

There are various algebraic and geometric characteristics that the KdV equation possesses. More significantly, a lot of physically important solutions to the KdV equation can be presented explicitly through a simple, specific form, called the *Hirota bilinear form*. More complicated types of solutions leading to N-solitons are presented in ► [Korteweg–de Vries Equation \(KdV\), History, Exact N-Soliton Solutions and Further Properties of the](#).

It was suitable to search for an easier method to solve this famous and well-known nonlinear partial differential equation (KdV). Many semi-analytical methods have been developed and used to solve the KdV equation. The most popular and famous semi-analytical method that was introduced to solve this nonlinear partial differential equation (KdV), was the *Adomian Decomposition Method*. ► [Adomian Decomposition Method Applied to Non-linear Evolution Equations in Soliton Theory](#).

Other semi-analytical methods like the variational iteration method (VIM), homotopy analysis method (HAM), and homotopy perturbation method (HPM) are usable to solve these types of NLPDE such as the Korteweg–de Vries equation and even the modified Korteweg–de Vries equation (mKdV). ► [Korteweg–de Vries Equation \(KdV\) and Modified Korteweg–de Vries Equations \(mKdV\), Semi-analytical Methods for Solving the](#).

It is indispensable to introduce other methods to illustrate the solution of this NLPDE. The numerical tech-

nique is a powerful tool which is always needed to present and compare the results in the easiest manner. There exist different numerical tools for solving the KdV and mKdV equations, and even for any evolution equations ► [Korteweg–de Vries Equation \(KdV\), Some Numerical Methods for Solving the](#).

In hydrodynamical problems, the motion of incompressible inviscid fluid subject to a constant vertical gravitational force (where the fluid is bounded below by an impermeable bottom and above by a free surface) with some physical assumptions lead to the shallow water model. The shallow water wave equations are usually used in oceanography and atmospheric science. The shallow water approximation theory leads to the KdV equation. This means that the soliton solution as well as the solitary waves are the straight forward solutions for many physical problems. ► [Shallow Water Waves and Solitary Waves](#).

The study of the water waves in the Pacific Ocean leads directly to a special type of long gravity waves named *tsunamis*. This word is a Japanese one, which means harbor's wave.

A tsunami is essentially a long wavelength water wave train or a series of waves generated in a body of water (mostly in oceans) that vertically displaces the water column. Earthquakes, landslides, volcanic eruptions, nuclear explosions and the impact of cosmic bodies can generate tsunamis. Tsunamis, as they approach coastlines, can rise enormously and savagely attack and inundate to cause huge damage to properties and cost thousands of lives.

The theoretical study of waves in oceanography is very complicated. The simplest mathematical model includes the KdV equation. This KdV represents the governing equation that lead to tsunamis. ► [Solitons, Tsunamis and Oceanographical Applications of](#).

The rapid change in water density with depth is called *Pycnocline*. This happens in open or closed seas, as well as in oceans. In the frame of shallow water approximation theory that governs the oceans, and due to the pycnocline the nonlinear internal gravity waves are presented. They arise from perturbations to hydrostatic equilibrium, where balance is maintained between the force of gravity and the buoyant restoring force. An illustrative study of nonlinear waves (that was generated inside a stratified fluid occupying a semi infinite channel of finite and constant depth by a wave maker situated in motion at the finite extremity of the channel) is presented in this section. ► [Non-linear Internal Waves](#).

The soliton perturbation theory is used to study solitons that are governed by the various nonlinear equations in the presence of the perturbation terms. ► [Soliton Perturbation](#).

A *compacton* is a special solitary traveling wave that, unlike a soliton, does not have exponential tails.

Another expression and terminology called *compacton-like soliton* is a special wave solution which can be expressed by the squares of sinusoidal or cosinusoidal functions.

Soliton and compacton are two kinds of nonlinear waves. They play an indispensable and vital role in all branches of science and technology, and are used as constructive elements to formulate the complex dynamical behavior of wave systems throughout science: from hydrodynamics to nonlinear optics, from plasmas to shock waves, from tornados to the Great Red Spot of Jupiter, and from tsunamis to turbulence. More recently, soliton and compacton have been of key importance in the quantum fields and nanotechnology especially in nano-hydrodynamics.

► **Solitons and Compactons.**

## Solitons, Tsunamis and Oceanographical Applications of

M. LAKSHMANAN

Center for Nonlinear Dynamics, Bharathidasan University, Tiruchirapalli, India

### Article Outline

Glossary

Definition of the Subject

Introduction

Shallow Water Waves and KdV Type Equations

Deep Water Waves and NLS Type Equations

Tsunamis as Solitons

Internal Solitons

Rossby Solitons

Bore Solitons

Future Directions

Bibliography

### Glossary

**Soliton** A class of nonlinear dispersive wave equations in (1+1) dimensions having a delicate balance between dispersion and nonlinearity admit localized solitary waves which under interaction retain their shapes and speeds asymptotically. Such waves are called solitons because of their particle like elastic collision property. The systems include Korteweg–de Vries, nonlinear Schrödinger, sine-Gordon and other nonlinear evolution equations. Certain (2+1) dimensional generalizations of these systems also admit soliton solutions of

different types (plane solitons, algebraically decaying lump solitons and exponentially decaying dromions).

**Shallow and deep water waves** Considering surface gravity waves in an ocean of depth  $h$ , they are called shallow-water waves if  $h \ll \lambda$ , where  $\lambda$  is the wavelength (or from a practical point of view if  $h < 0.07\lambda$ ). In the linearized case, for shallow water waves the phase speed  $c = \sqrt{gh}$ , where  $g$  is the acceleration due to gravity. Water waves are classified as deep (practically) if  $h > 0.28\lambda$  and the corresponding wave speed is given by  $c = \sqrt{g/k}$ ,  $k = \frac{2\pi}{\lambda}$ .

**Tsunami** Tsunami is essentially a long wavelength water wave train, or a series of waves, generated in a body of water (mostly in oceans) that vertically displaces the water column. Earthquakes, landslides, volcanic eruptions, nuclear explosions and impact of cosmic bodies can generate tsunamis. Propagation of tsunamis is in many cases in the form of shallow water waves and sometimes can be of the form of solitary waves/solitons. Tsunamis as they approach coastlines can rise enormously and savagely attack and inundate to cause devastating damage to life and property.

**Internal solitons** Gravity waves can exist not only as surface waves but also as waves at the interface between two fluids of different density. While solitons were first recognized on the surface of water, the commonest ones in oceans actually happen underneath, as internal oceanic waves propagating on the pycnocline (the interface between density layers). Such waves occur in many seas around the globe, prominent among them being the Andaman and Sulu seas.

**Rossby solitons** Rossby waves are typical examples of quasigeostrophic dynamical response of rotating fluid systems, where long waves between layers of the atmosphere as in the case of the Great Red Spot of Jupiter or in the barotropic atmosphere are formed and may be associated with solitonic structures.

**Bore solitons** The classic bore (also called mascaret, poroca and aeger) arises generally in funnel shaped estuaries that amplify incoming tides, tsunamis or storm surges, the rapid rise propagating upstream against the flow of the river feeding the estuary. The profile depends on the Froude number, a dimensionless ratio of inertial and gravitational effects. Slower bores can take on oscillatory profile with a leading dispersive shock-wave followed by a train of solitons.

### Definition of the Subject

Surface and internal gravity waves arising in various oceanographic conditions are natural sources where one

can identify/observe the generation, formation and propagation of solitary waves and solitons. Unlike the standard progressive waves of linear dispersive type, solitary waves are localized structures with long wavelengths and finite energies and propagate without change of speed or form and are patently nonlinear entities. The earliest scientifically recorded observation of a solitary wave was made by John Scott Russel in August 1834 in the Union Canal connecting the Scottish cities of Glasgow and Edinburgh. The theoretical formulation of the underlying phenomenon was provided by Korteweg and de Vries in 1895 who deduced the now famous Korteweg–de Vries (KdV) equation admitting solitary wave solutions. With the insightful numerical and analytical investigations of Martin Kruskal and coworkers in the 1960s the KdV solitary waves have been shown to possess the remarkable property that under collision they pass through each other without change of shape or speed except for a phase shift and so they are solitons. Since then a large class of soliton possessing nonlinear dispersive wave equations such as the sine-Gordon, modified KdV (MKdV) and nonlinear Schrödinger (NLS) equations occurring in a wide range of physical phenomena have been identified.

Several important oceanographic phenomena which correspond to nonlinear shallow water wave or deep water wave propagation have been identified/interpreted in terms of soliton propagation. These include tsunamis, especially earthquake induced ones like the 1960 Chilean or 2004 Indian Ocean earthquakes, internal solitary waves arising in stratified stable fluids such as the ones observed in Andaman or Sulu seas, Rossby waves including the Giant Red Spot of Jupiter and tidal bores occurring in estuaries of rivers. Detailed observations/laboratory experiments and theoretical formulations based on water wave equations resulting in the nonlinear evolution equations including KdV, Benjamin–Ono, Intermediate Long Wave (ILW), Kadomtsev–Petviashvili (KP), NLS, Davey–Stewartson (DS) and other equations clearly establish the relevance of soliton description in such oceanographic events.

## Introduction

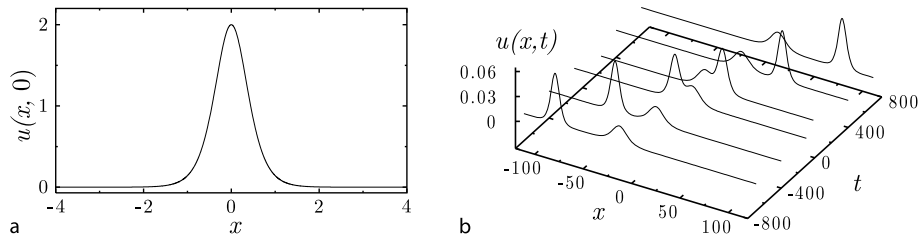
Historically, the remarkable observation of John Scott Russel [1,2] of the solitary wave in the Union Canal connecting the cities of Edinburgh and Glasgow in the month of August 1834 may be considered as the precursor to the realization of solitons in many oceanographic phenomena. While riding on a horse back and observing the motion of boat drawn by a pair of horses which suddenly stopped but not so the mass of water it had set in motion, the wave

(which he called the ‘Great Wave of Translation’) in the form of large solitary heap of water surged forward and travelled a long distance without *change of form or diminution of speed*. The wave observed by Scott Russel is nothing but a solitary wave having a remarkable staying power and a patently nonlinear entity. Korteweg and de Vries in 1895, starting from the basic equations of hydrodynamics and considering unidirectional shallow water wave propagation in rectangular channels, deduced [3] the now ubiquitous KdV equation as the underlying nonlinear evolution equation. It is a third order nonlinear partial differential equation in (1+1) dimensions with a delicate balance between dispersion and nonlinearity. It admits elliptic function cnoidal wave solutions and in a limiting form exact solitary wave solution of the type observed by John Scott Russel thereby vindicating his observations and putting to rest all the controversies surrounding them.

It was the many faceted numerical and analytical study of Martin Kruskal and coworkers [4,5] which firmly established by 1967 that the KdV solitary waves have the further remarkable feature that they are solitons having elastic collision property (Fig. 1).

It was proved decisively that the KdV solitary waves on collision pass through each other except for a finite phase shift, thereby retaining their forms and speeds asymptotically as in the case of particle like elastic collisions. The inverse scattering transform (IST) formalism developed for this purpose clearly shows that the KdV equation is a completely integrable infinite dimensional nonlinear Hamiltonian system and that it admits multisoliton solutions [6,7,8,9]. Since then a large class of nonlinear dispersive wave equations such as the sine-Gordon (s-G), modified Korteweg–de Vries (MKdV), NLS, etc. equations in (1+1) dimensions modeling varied physical phenomena have also been shown to be completely integrable soliton systems [6,7,8,9]. Interesting (2+1) dimensional versions of these systems such as Kadomtsev–Petviashvili (KP), Davey–Stewartson (DS) and Nizhnik–Novikov–Veselov (NNV) equations have also been shown to be integrable systems admitting basic nonlinear excitations such as line (plane) solitons, algebraically decaying lump solitons and exponentially localized dromion solutions [7,8].

It should be noted that not all solitary waves are solitons while the converse is always true. An example of a solitary wave which is not a soliton is the one which occurs in double-well  $\lambda\phi^4$  wave equation which radiates energy on collision with another such wave. However even such solitary waves having finite energies are sometimes referred to as solitons in the condensed matter, particle physics and fluid dynamics literature because of their localized structure.



**Solitons, Tsunamis and Oceanographical Applications of, Figure 1**

**KdV equation: a One soliton solution (solitary wave) at a fixed time, say  $t = 0$ , b Two soliton solution (depicting elastic collision)**

As noted above solitary waves and solitons are abundant in oceanographic phenomena, especially involving shallow water wave and deep water wave propagations. However, these events are large scale phenomena very often difficult to measure experimentally, rely mostly on satellite and other indirect observations and so controversies and differing interpretations do exist in ascribing exact solitonic or solitary wave properties to these phenomena. Yet it is generally realized that many of these events are closely identifiable with solitonic structures. Some of these observations deserve special attention.

### Tsunamis

When large scale earthquakes, especially of magnitude 8.0 in Richter scale and above, occur in seabeds at appropriate geological faults tsunamis are generated and can propagate over large distances as small amplitude and long wavelength structures when shallowness condition is satisfied. Though the tsunamis are hardly felt in the mid-sea, they take monstrous structures when they approach land masses due to conservation of energy. The powerful Chilean earthquake of 1960 [10] led to tsunamis which propagated for almost fifteen hours before striking Hawaii islands and a further seven hours later they struck the Japanese islands of Honshu and Hokkaido. More recently the devastating Sumatra–Andaman earthquake of 2004 in the Indian Ocean generated tsunamis which not only struck the coastlines of Asian countries including Indonesia, Thailand, India and Srilanka but propagated as far as Somalia and Kenya in Africa, killing more than a quarter million people. There is very good sense in ascribing soliton description to such tsunamis.

### Internal Solitons

Peculiar striations, visible on satellite photographs of the surface of the Andaman and Sulu seas in the far east (and in many other oceans around the globe), have been interpreted as secondary phenomena accompanying the pas-

sage of “internal solitons”. These are solitary wavelike distortions of the boundary layer between the warm upper layer of sea water and cold lower depths. These internal solitons are travelling ridges of warm water, extending hundreds of meters down below the thermal boundary, and carry enormous energy. Osborne and Burch [11] investigated the underwater currents which were experienced by an oil rig in the Andaman sea, which was drilling at a depth of 3600 ft. One drilling rig was apparently spun through ninety degrees and moved one hundred feet by the passage of a soliton below.

### Rossby Waves and Solitons

Rossby waves [12] are long waves between layers of the atmosphere, created by the rotation of the planet. In particular, in the atmosphere of a rotating planet, a fluid particle is endowed with a certain rotation rate, determined by its latitude. Consequently its motion in the north-south direction is constrained by the conservation of angular momentum as in the case of internal waves where gravity inhibits the vertical motion of a density stratified fluid. There is an analogy between internal waves and Rossby waves under suitable conditions. The KdV equation has been proposed as a model for the evolution of Rossby solitons [13] and NLS equation for the evolution of Rossby wave packets. The Great Red Spot of the planet of Jupiter is often associated with a Rossby soliton.

### Bore Solitons

Rivers which flow to the open oceans are very often affected by tidal effects, tsunamis or storm surges. Tidal motions generate intensive water flows which can propagate upstream on tens of kilometers in the form of step-wise perturbation (hydraulic jumps) analogous to shock waves in acoustics. This phenomenon is known as a bore or mascaret (in French). Examples of such bores include the tidal bore of Seine river in France, Hooghly bore of the Ganges in India, the Amazon river bore in Brazil and Hangzhou

bore in China. Bore disintegration into solitons is a possible phenomenon in such bores [14].

Besides these there are other possible oceanographic phenomena such as capillary wave solitons [15], resonant three-wave or four wave interaction solitons [16], etc., where also soliton picture is useful.

In this article we will first point out how in shallow channels the long wavelength wave propagation is described by KdV equation and its generalizations (Sect. “Shallow Water Waves and KdV Type Equations”). Then we will briefly point out how NLS family of equations arise naturally in the description of deep water waves (Sect. “Deep Water Waves and NLS Type Equations”). Based on these details, we will point how the soliton picture plays a very important role in the understanding of tsunami propagation (Sect. “Tsunamis as Solitons”), generation of internal solitons (Sect. “Internal Solitons”), formation of Rossby solitons (Sect. “Rossby Solitons”) and disintegration of bores into solitons (Sect. “Bore Solitons”).

### Shallow Water Waves and KdV Type Equations

Kortweg and de Vries [3] considered the wave phenomenon underlying the observations of Scott Russel from first principles of fluid dynamics and deduced the KdV equation to describe the unidirectional shallow water wave propagation in one dimension.

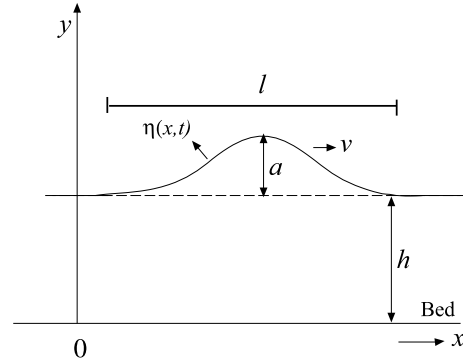
Consider the one-dimensional ( $x$ -direction) wave motion of an incompressible and inviscid fluid (water) in a shallow channel of height  $h$ , and of sufficient width with uniform cross-section leading to the formation of a solitary wave propagating under gravity. The effect of surface tension is assumed to be negligible. Let the length of the wave be  $l$  and the maximum value of its amplitude,  $\eta(x, t)$ , above the horizontal surface be  $a$  (see Fig. 2).

Then assuming  $a \ll h$  (shallow water) and  $h \ll l$  (long waves), one can introduce two natural small parameters into the problem  $\epsilon = a/h$  and  $\delta = h/l$ . Then the analysis proceeds as follows [3,6].

#### Equation of Motion: KdV Equation

The fluid motion can be described by the velocity vector  $\mathbf{V}(x, y, t) = u(x, y, t)\mathbf{i} + v(x, y, t)\mathbf{j}$ , where  $\mathbf{i}$  and  $\mathbf{j}$  are the unit vectors along the horizontal and vertical directions, respectively. As the motion is irrotational, we have  $\nabla \times \mathbf{V} = 0$ . Consequently, we can introduce the velocity potential  $\phi(x, y, t)$  by the relation  $\mathbf{V} = \nabla\phi$ .

**Conservation of Density** The system obviously admits the conservation law for the mass density  $\rho(x, y, t)$  of the



Solitons, Tsunamis and Oceanographical Applications of, Figure 2

One-dimensional wave motion in a shallow channel

fluid,  $d\rho/dt = \rho_t + \nabla \cdot (\rho\mathbf{V}) = 0$ . As  $\rho$  is a constant, we have  $\nabla \cdot \mathbf{V} = 0$ . Consequently  $\phi$  obeys the Laplace equation

$$\nabla^2 \phi(x, y, t) = 0. \quad (1)$$

**Euler's Equation** As the density of the fluid  $\rho = \rho_0$  is constant, using Newton's law for the rate of change of momentum, we can write  $d\mathbf{V}/dt = \partial\mathbf{V}/\partial t + (\mathbf{V} \cdot \nabla)\mathbf{V} = -\frac{1}{\rho_0}\nabla p - g\mathbf{j}$ , where  $p = p(x, y, t)$  is the pressure at the point  $(x, y)$  and  $g$  is the acceleration due to gravity, which is acting vertically downwards (here  $\mathbf{j}$  is the unit vector along the vertical direction). Since  $\mathbf{V} = \nabla\phi$  we obtain (after one integration)

$$\phi_t + \frac{1}{2}(\nabla\phi)^2 + \frac{p}{\rho_0} + gy = 0. \quad (2)$$

**Boundary Conditions** The above two Eqs. (1) and (2) for the velocity potential  $\phi(x, y, t)$  of the fluid have to be supplemented by appropriate boundary conditions, by taking into account the fact (see Fig. 2) that (a) the horizontal bed at  $y = 0$  is hard and (b) the upper boundary  $y = y(x, t)$  is a free surface. As a result

(a) the vertical velocity at  $y = 0$  vanishes,  $v(x, 0, t) = 0$ , which implies

$$\phi_y(x, 0, t) = 0. \quad (3)$$

(b) As the upper boundary is free, let us specify it by  $y = h + \eta(x, t)$  (see Fig. 2). Then at the point  $x = x_1, y = y_1 \equiv y(x, t)$ , we can write  $\frac{dy_1}{dt} = \frac{\partial\eta}{\partial t} + \frac{\partial\eta}{\partial x} \frac{dx_1}{dt} = \eta_t + \eta_x u_1 = v_1$ . Since  $v_1 = \phi_{1y}, u_1 = \phi_{1x}$ , we obtain

$$\phi_{1y} = \eta_t + \eta_x \phi_{1x}. \quad (4)$$



(c) Similarly at  $y = y_1$ , the pressure  $p_1 = 0$ . Then from (2), it follows that

$$u_{1t} + u_1 u_{1x} + v_1 v_{1x} + g \eta_x = 0. \quad (5)$$

Thus the motion of the surface of water wave is essentially specified by the Laplace Eq. (1) and Euler Eq. (2) along with one fixed boundary condition (3) and two *variable nonlinear* boundary conditions (4) and (5). One has to then solve the Laplace equation subject to these boundary conditions.

**Taylor Expansion of  $\phi(x, y, t)$  in  $y$**  Making use of the fact  $\delta = h/l \ll 1$ ,  $h \ll l$ , we assume  $y (= h + \eta(x, t))$  to be small to introduce the Taylor expansion

$$\phi(x, y, t) = \sum_{n=0}^{\infty} y^n \phi_n(x, t). \quad (6)$$

Substituting the above series for  $\phi$  into the Laplace Eq. (1), solving recursively for  $\phi_n(x, t)$ 's and making use of the boundary condition (4),  $\phi_y(x, 0, t) = 0$ , one can show that

$$u_1 = \phi_{1x} = f - \frac{1}{2} y_1^2 f_{xx} + \text{higher order in } y_1, \quad (7)$$

$$v_1 = \phi_{1y} = -y_1 f_x + \frac{1}{6} y_1^3 f_{xxx} + \text{higher order in } y_1, \quad (8)$$

where  $f = \partial \phi_0 / \partial x$ . We can then substitute these expressions into the nonlinear boundary conditions (4) and (5) to obtain equations for  $f$  and  $\eta$ .

**Introduction of Small Parameters  $\epsilon$  and  $\delta$**  So far the analysis has not taken into account fully the shallow nature of the channel ( $a/h = \epsilon \ll 1$ ) and the solitary nature of the wave ( $a/l = a/h \cdot h/l = \epsilon \delta \ll 1$ ,  $\epsilon \ll 1$ ,  $\delta \ll 1$ ), which are essential to realize the Scott Russel phenomenon. For this purpose one can stretch the independent and dependent variables in the above equations through appropriate scale changes, but retaining the overall form of the equations. To realize this one can introduce the natural scale changes

$$x = l x', \quad \eta = a \eta', \quad t = \frac{l}{c_0} t', \quad (9)$$

where  $c_0$  is a parameter to be determined. Then in order to retain the form of (7) and (8) we require

$$u_1 = \epsilon c_0 u'_1, \quad v_1 = \epsilon \delta c_0 v'_1, \quad f = \epsilon c_0 f', \quad (10)$$

$$y_1 = h + \eta(x, t) = h (1 + \epsilon \eta' (x', t')) .$$

Then

$$u'_1 = f' - \frac{1}{2} \delta^2 (1 + \epsilon \eta')^2 f'_{x'x'} = f' - \frac{1}{2} \delta^2 f'_{x'x'}, \quad (11)$$

where we have omitted terms proportional to  $\delta^2 \epsilon$  as small compared to terms of the order  $\delta^2$ . Similarly from (8), we obtain

$$v'_1 = - (1 + \epsilon \eta') f'_{x'} + \frac{1}{6} \delta^2 f'_{x'x't'}. \quad (12)$$

Now considering the nonlinear boundary condition (4) in the form  $v_1 = \eta_t + \eta_x u_1$ , it can be rewritten as

$$\eta'_{t'} + f'_{x'} + \epsilon \eta' f'_{x'} + \epsilon f' \eta'_{x'} - \frac{1}{6} \delta^2 f'_{x'x't'} = 0. \quad (13)$$

Similarly considering the other boundary condition (5) and making use of the above transformations, it can be rewritten, after neglecting terms of the order  $\epsilon^2 \delta^2$ , as

$$f'_{t'} + \epsilon f' f'_{x'} + \frac{g a}{\epsilon c_0^2} \eta'_{x'} - \frac{1}{2} \delta^2 f'_{x'x't'} = 0. \quad (14)$$

Now choosing the arbitrary parameter  $c_0$  as  $c_0^2 = gh$  so that  $\eta'_{x'}$  term is of order unity, (14) becomes

$$f'_{t'} + \eta'_{x'} + \epsilon f' f'_{x'} - \frac{1}{2} \delta^2 f'_{x'x't'} = 0. \quad (15)$$

(Note that  $c_0 = \sqrt{gh}$  is nothing but the speed of the water wave in the linearized limit). Omitting the primes for convenience, the evolution equation for the amplitude of the wave and the function related to the velocity potential reads

$$\eta_t + f_x + \epsilon \eta f_x + \epsilon f \eta_x - \frac{1}{6} \delta^2 f_{xxx} = 0, \quad (16)$$

$$f_t + \eta_x + \epsilon f f_x - \frac{1}{2} \delta^2 f_{xxt} = 0. \quad (17)$$

Note that the small parameters  $\epsilon$  and  $\delta^2$  have occurred in a natural way in (16), (17).

**Perturbation Analysis** Since the parameters  $\epsilon$  and  $\delta^2$  are small in (16), (17), we can make a perturbation expansion of  $f$  in these parameters:

$$f = f^{(0)} + \epsilon f^{(1)} + \delta^2 f^{(2)} + \text{higher order terms}, \quad (18)$$

where  $f^{(i)}$ ,  $i = 0, 1, 2, \dots$  are functions of  $\eta$  and its spatial derivatives. Note that the above perturbation expansion is an asymptotic expansion. Substituting this into Eqs. (16) and (17) and regrouping and comparing different powers proportional to  $(\epsilon, \delta^2)$  and solving them successively one

can obtain (see for example [6] for further details) in a self consistent way,

$$f^{(0)} = \eta, \quad f^{(1)} = -\frac{1}{4}\eta^2, \quad f^{(2)} = \frac{1}{3}\eta_{xx}. \quad (19)$$

Using these expressions into (18) and substituting it in (16) and (17), we ultimately obtain the KdV equation in the form

$$\eta_t + \eta_x + \frac{3}{2}\epsilon\eta\eta_x + \frac{\delta^2}{6}\eta_{xxx} = 0, \quad (20)$$

describing the unidirectional propagation of long wavelength shallow water waves.

### The Standard (Contemporary) Form of KdV Equation

Finally, changing to a moving frame of reference,  $\xi = x - t$ ,  $\tau = t$ , and introducing the new variables  $u = (3\epsilon/2\delta^2)\eta$ ,  $\tau' = (6/\delta^2)\tau$  and redefining the variables  $\tau'$  as  $t$  and  $\xi$  as  $x$ , we finally arrive at the ubiquitous form of the KdV equation as

$$u_t + 6uu_x + u_{xxx} = 0. \quad (21)$$

The Korteweg–de Vries Eq. (21) admits cnoidal wave solution and in the limiting case solitary wave solution as well. This form can be easily obtained [6,17] by looking for a wave solution of the form  $u = 2f(\xi)$ ,  $\xi = (x - ct)$ , and reducing the KdV equation into a third order nonlinear ordinary differential equation of the form

$$-c \frac{\partial f}{\partial \xi} + 12f \frac{\partial f}{\partial \xi} + \frac{\partial^3 f}{\partial \xi^3} = 0. \quad (22)$$

Integrating Eq. (22) twice and rearranging, we can obtain

$$\left(\frac{\partial f}{\partial \xi}\right)^2 = -4f^3 + cf^2 - 2df - 2b \equiv P(f), \quad (23)$$

where  $b$  and  $d$  are integration constants. Calling the three real roots of the cubic equation  $P(f) = 0$  as  $\alpha_1$ ,  $\alpha_2$  and  $\alpha_3$  such that

$$\left(\frac{\partial f}{\partial \xi}\right)^2 = -4(f - \alpha_1)(f - \alpha_2)(f - \alpha_3), \quad (24)$$

the solution can be expressed in terms of the Jacobian elliptic function as

$$f(\xi) = f(x - ct) = \frac{\alpha_3 - (\alpha_3 - \alpha_2)\text{sn}^2}{\left[\sqrt{\alpha_3 - \alpha_1}(x - ct), m\right]}, \quad (25a)$$

where

$$(\alpha_1 + \alpha_2 + \alpha_3) = \frac{c}{4}, \quad m^2 = \frac{\alpha_3 - \alpha_2}{\alpha_3 - \alpha_1}, \quad \delta = \text{constant}. \quad (25b)$$

Here  $m$  is the modulus parameter of the Jacobian elliptic function. Eq. (25a) represents in fact the so called cnoidal wave (because of its elliptic function form). In the limiting case  $m = 1$ , the form (25a) reduces to

$$f = \alpha_2 + (\alpha_3 - \alpha_2) \text{sech}^2 \left[ \sqrt{\alpha_3 - \alpha_1} (x - ct) \right]. \quad (26)$$

Choosing now  $\alpha_1 = 0$ ,  $\alpha_2 = 0$ , and using (25b) we have

$$f = \frac{c}{4} \text{sech}^2 \left[ \frac{\sqrt{c}}{2} (x - ct) \right]. \quad (27)$$

Then the solitary wave solution to the KdV Eq. (21) can be written in the form

$$u(x, t) = \frac{c}{2} \text{sech}^2 \frac{\sqrt{c}}{2} (x - ct + \delta), \quad \delta : \text{constant} \quad (28)$$

Note that the velocity of the solitary wave is directly proportional to the amplitude: larger the wave the higher is the speed. More importantly, the KdV solitary wave is a soliton: it retains its shape and speed upon collision with another solitary wave of different amplitude, except for a phase shift, see Fig. 1 [6,7,8]. In fact for an arbitrary initial condition, the solution of the Cauchy initial value problem consists of N-number of solitons of different amplitudes in the background of small amplitude dispersive waves. All these results ultimately lead to the result that the KdV equation is a completely integrable, infinite dimensional, nonlinear Hamiltonian system. It possesses [6,7,8]

- (i) a Lax pair of linear differential operators and is solvable through the so called inverse scattering transform (IST) method,
- (ii) infinite number of conservation laws and associated infinite number of involutive integrals of motion,
- (iii) N-soliton solution,
- (iv) Hirota bilinear form,
- (v) Hamiltonian structure

and a host of other interesting properties (see for example [6,7,8,9]).

### KdV Related Integrable and Nonintegrable NLEEs

Depending on the actual physical situation, the derivation of the shallow water wave equation can be suitably modified to obtain other forms of nonlinear dispersive wave

equations in (1+1) dimensions as well as in (2+1) dimensions relevant for the present context. Without going into the actual derivations, some of the important equations possessing solitary waves are listed below [6,7,8,9].

1. Boussinesq equation [7]

$$u_t + uu_x + g\eta_x - \frac{1}{3}h^2u_{txx} = 0, \quad (29a)$$

$$\eta_t + [u(h + \eta)]_x = 0 \quad (29b)$$

2. Benjamin–Bona–Mahoney (BBM) equation [18]

$$u_t + u_x + uu_x - u_{xxt} = 0 \quad (30)$$

3. Camassa–Holm equation [19]

$$u_t + 2\kappa u_x + 3uu_x - u_{xxt} = 3u_x u_{xx} + uu_{xxx} \quad (31)$$

4. Kadomtsev–Petviashvili (KP) equation [7]

$$(u_t + 6uu_x + u_{xxx})_x + 3\sigma^2 u_{yy} = 0 \quad (32)$$

$$(\sigma^2 = -1: \text{KP-I}, \sigma^2 = +1: \text{KP-II}).$$

In the derivation of the above equations, generally the bottom of water column or fluid bed is assumed to be flat. However in realistic situations the water depth varies as a function of the horizontal coordinates. In this situation, one often encounters inhomogeneous forms of the above wave equations. Typical example is the variable coefficient KdV equation [14]:

$$u_t + f(x, t)uu_x + g(x, t)u_{xxx} = 0, \quad (33)$$

where  $f$  and  $g$  are functions of  $x, t$ . More general forms can also be deduced depending upon the actual situations, see for example [14].

### Deep Water Waves and NLS Type Equations

Deep water waves are strongly dispersive in contrast to the weakly dispersive nature of the shallow water waves (in the linear limit). Various authors (see for details [8]) have shown that nonlinear Schrödinger family of equations models the evolution of a packet of surface waves in this case. There are several oceanographic situations where such waves can arise [8]:

(i) A localized storm at sea can generate a wide spectrum of waves, which then propagates away from the source region in all horizontal directions. If the propagat-

ing waves have small amplitudes and encounter no wind away from the source region, these waves can eventually sort themselves into nearly one-dimensional packets of nearly monochromatic waves. For appropriately chosen scales, the underlying evolution of each of these packets can be shown to satisfy the nonlinear Schrödinger equation and its generalizations in (2+1) dimensions.

(ii) Nearly monochromatic, nearly one-dimensional waves can cover a broad range of surface waves in the sea that results due to a steady wind of long duration and fetch. Then the generalization of the NLS equation in (2+1) dimensions can describe the waves that result in when the wind stops.

In all the above situations one looks for the solution of the equations of motion (1)-(5) but generalized in three dimensions which consists mainly in the form of a small amplitude, nearly monochromatic, nearly one-dimensional wave train. Assuming that this wave train travels in the  $x$ -direction with a mean wave number  $\kappa = (k, l)$  with a characteristic amplitude ' $a$ ' of the disturbance and a characteristic variation  $\delta k$  in  $k$ , one can deduce the NLS equation in (2+1) dimensions under the following conditions:

- (i) small amplitudes such that  $\hat{\epsilon} = \epsilon\delta \equiv \kappa a \ll 1$
- (ii) slowly varying modulations,  $\frac{\delta k}{\kappa} \ll 1$
- (iii) nearly one dimensional waves,  $\frac{|l|}{k} \ll 1$
- (iv) balance of all three effects,  $\frac{\delta k}{\kappa} = \frac{|l|}{k} \approx O(\hat{\epsilon})$
- (v) for finite and deep water waves,  $(kh)^2 \gg \hat{\epsilon}$

In the lowest order approximation (linear approximation) of water waves, the prediction is that a localized initial state will generally evolve into wave packets with a dominant wave number  $\kappa$  and corresponding frequency  $\omega$ , given by the dispersion relation  $\omega = (g\kappa + \sigma\kappa^3)\tanh\kappa h$ ,  $\kappa = |\kappa| = \sqrt{k^2 + l^2}$  within which each wave propagates with the phase speed  $c = \omega/k$ , while the envelope propagates with the group velocity  $c_g = d\omega/d\kappa$ . After a sufficiently long time the wave packet tends to disperse around the dominant wave number.

This tendency for dispersion can be offset by cumulative nonlinear effects. In the absence of surface tension, the outcome for unidirectional waves can be shown to be describable by the NLS equation. If the surface wave in the lowest order is  $\phi \approx A \exp i(kx - \omega t) + c.c.$ , where  $\phi$  is the velocity potential, then to leading order the wave amplitude evolves as

$$i(A_t + c_g A_x) + \frac{1}{2}\lambda A_{xx} + \mu |A|^2 A = 0. \quad (34)$$

The coefficients here are given by

$$\begin{aligned}\lambda &= \frac{\partial^2 \omega}{\partial k^2}, \\ \mu &= -\frac{\omega k^2}{16S^4}(8C^2S^2 + g - 2T^2) \\ &\quad + \frac{\omega}{8C^2S^2} \frac{(2\omega C^2 + kc_g)^2}{(gh - c_g^2)},\end{aligned}\quad (35)$$

where  $C = \cosh(kh)$ ,  $S = \sinh(kh)$ ,  $T = \frac{S}{C}$ . Equation (34) has been obtained originally by Zakharov in 1968 for deep water waves [20] and by Hasimoto and Ono for waves of finite depth in 1972 [21].

The NLS equation is also a soliton possessing integrable system and is solvable by the IST method [6,7,8,9]. For  $\lambda > 0$  (focusing case), the envelope solitary (soliton) wave solution (also called bright solitons in the optical physics context) is given by

$$A(x, t) = a \operatorname{sech} \gamma(x - c_g t) \exp(-i\Omega t), \quad (36)$$

where  $\mu a^2 = \lambda \gamma^2$ ,  $\Omega = +\frac{1}{2}\mu a^2$ .

When the effects of modulation in the transverse  $y$ -direction are taken into account, so that the wave amplitude is now given by  $A(x, y, t)$ , the NLS equation is replaced by the Benney–Roskes system [22] also popularly known as the Davey–Stewartson equations [23],

$$i(A_t + c_g A_x) + \frac{1}{2}\lambda A_{xx} + \frac{1}{2}\delta A_{yy} + \mu |A|^2 A + UA = 0, \quad (37a)$$

$$\alpha U_{xx} + U_{yy} + \beta(|A|^2)_{yy} = 0, \quad (37b)$$

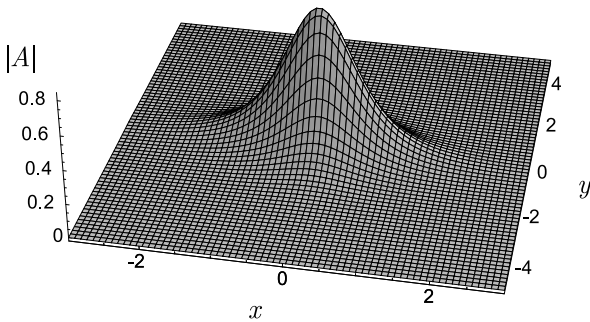
where  $\delta = \frac{c_g}{k}$ ,  $\alpha = 1 - \left(\frac{c_g^2}{gh}\right)$ ,  $gh\beta = \frac{\omega}{8C^2S^2}(2\omega C^2 + kc_g)^2$ . Here  $U(x, y, t)$  is the wave induced mean flow. In

the deep water wave limit,  $kh \rightarrow \infty$ , and  $U \rightarrow 0$ ,  $\beta \rightarrow 0$  and one has the nonintegrable (2+1) dimensional NLS equation. On the other hand in the shallow water limit, one has the integrable Davey–Stewartson (DS) equations. For details see [7,8]. The DS-I equation admits algebraically decaying lump solitons and exponentially decaying dromions [24] besides the standard line solitons for appropriate choices of parameters. A typical dromion solution of DS-I equation is shown in Fig. 3.

### Tsunamis as Solitons

The term ‘tsunami’ (tsu:harbour, nami:wave in Japanese) which was perhaps an unknown word even for scientists in countries such as India, Srilanka, Thailand, etc. till recently has become a house-hold word since that fateful morning of December 26, 2004. When a powerful earthquake of magnitude 9.1–9.3 on the Richter scale, epicentered off the coast of Sumatra, Indonesia, struck at 07:58:53, local time, described as the 2004 Indian Ocean earthquake or Sumatra–Andaman earthquake (Fig. 4), it triggered a series of devastating tsunamis as high as 30 meters that spread throughout the Indian Ocean killing about 275,000 people and inundating coastal communities across South and Southeast Asia, including parts of Indonesia, Srilanka, India and Thailand and even reaching as far as the east coast of Africa [25]. The catastrophe is considered to be one of the deadliest disasters in modern history.

Since this earthquake and consequent tsunamis, several other earthquakes of smaller and larger magnitudes keep occurring off the coast of Indonesia. Even as late as July 17, 2006 an earthquake of magnitude 7.7 on the Richter scale struck off the town of Pandering at 15.19 lo-



**Solitons, Tsunamis and Oceanographical Applications of, Figure 3**

Exponentially localized dromion solution of the Davey–Stewartson equation at a fixed time ( $t = 0$ ) for suitable choice of parameters



**Solitons, Tsunamis and Oceanographical Applications of, Figure 4**

26 December 2004 Indian Ocean tsunami (adapted from the website [www.blogaid.org.uk](http://www.blogaid.org.uk) with the courtesy of Andy Budd)

cal time and set off a tsunami of 2m high which had killed more than 300 people.

These tsunamis, which can become monstrous tidal waves when they approach coastline, are essentially triggered due to the sudden vertical rise of the seabed by several meters (when earthquake occurs) which displaces massive volume of water. The tsunamis behave very differently in deep water than in shallow water as pointed out below. By no means the tsunami of 2004 and later ones are exceptional; More than two hundred tsunamis have been recorded in scientific literature since ancient times. The most notable earlier one is the tsunami triggered by the powerful earthquake (9.6 magnitude) off southern Chile on May 22, 1960 [10] which traveled almost 22 hours before striking Japanese islands.

It is clear from the above events that the tsunami waves are fairly permanent and powerful ones, having the capacity to travel extraordinary distances without practically diminishing in size or speed. In this sense they seem to have considerable resemblance to shallow water nonlinear dispersive waves of KdV type, particularly solitary waves and solitons. It is then conceivable that tsunami dynamics has close connection with soliton dynamics.

### Basics of Tsunami Waves

As noted above tsunami waves of the type described above are essentially triggered by massive earthquakes which lead to vertical displacement of a large volume of water. Other possible reasons also exist for the formation and propagation of tsunami waves: underwater nuclear explosion, larger meteorites falling into the sea, volcano explosions, rockslides, etc. But the most predominant cause of tsunamis appear to be large earthquakes as in the case of the Sumatra–Andaman earthquake of 2004. Then there are three major aspects associated with the tsunami dynamics [26]:

1. Generation of tsunamis
2. Propagation of tsunamis
3. Tsunami run up and inundation

There exist rather successful models to approach the generation aspects of tsunamis when they occur due to the earthquakes [27]. Using the available seismic data it is possible to reconstruct the permanent deformation of the sea bottom due to earthquakes and simple models have been developed (see for example, [28]). Similarly the tsunami run up and inundation problems [29] are extremely complex and they require detailed critical study from a practical point of view in order to save structures and lives when a tsunami strikes.

However, here we will be more concerned with the propagation of tsunami waves and their possible relation to wave propagation associated with nonlinear dispersive waves in shallow waters. In order to appreciate such a possible connection, we first look at the typical characteristic properties of tsunami waves as in the case of 2004 Indian Ocean tsunami waves or 1960 Chilean tsunamis.

### The Indian Ocean Tsunami of 2004

Considering the Indian Ocean 2004 tsunami, satellite observations after a couple of hours after the earthquake establish an amplitude of approximately 60 cms in the open ocean for the waves. The estimated typical wavelength is about 200 kms [30]. The maximum water depth  $h$  is between 1 and 4 kms. Consequently, one can identify in an average sense the following small parameters ( $\epsilon$  and  $\delta^2$ ) of roughly equal magnitude:

$$\epsilon = \frac{a}{h} \approx 10^{-4} \ll 1, \quad \delta^2 = \frac{h^2}{l^2} \approx 10^{-4} \ll 1 \quad (38)$$

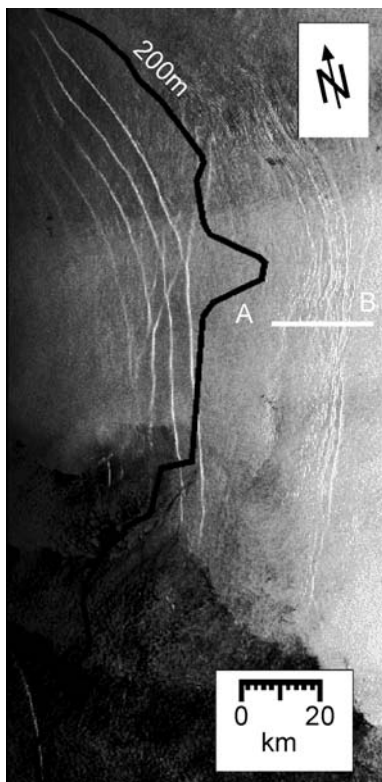
As a consequence, it is possible that a nonlinear shallow water wave theory where dispersion (KdV equation) also plays an important role (as discussed in Sect. “Shallow Water Waves and KdV Type Equations”) has considerable relevance [26]. However, we also wish to point out here that there are other points of view: Constantin and Johnson [31] estimate  $\epsilon \approx 0.002$  and  $\delta \approx 0.04$  and conclude that for both nonlinearity and dispersion to become significant the quantity  $\delta\epsilon^{-3/2} \times \text{wavelength}$  estimated as 90,000 kms is too large and shallow water equations with variable depth (without dispersion) should be used. However, it appears that these estimates can vary over a rather wide range and with suitable estimates it is possible that the range of 10,000–20,000 kms could be also possible ranges and hence taking into account the fact that both the Indian Ocean 2004 and Chilean 1960 tsunamis have traveled over 10 hours or more (in certain directions) before encountering land mass appears to allow for the possibility of nonlinear dispersive waves as relevant features for the phenomena. Segur [32] has argued that in the 2004 tsunamis, the propagation distances from the epicenter of the earthquake to India, Srilanka, or Thailand were too short for KdV dynamics to develop. In the same way one can argue that the waves that hit Somalia and Kenya in the east coast of Africa (or as in the case of Chilean earthquake see also [32]) have traveled sufficiently long distance for KdV dynamics to become important [33]. In any case one can conclude that at least for long distance tsunami propagation solitary wave and soliton picture of KdV like equations become very relevant.



## Internal Solitons

For a long time seafarers passing through the Strait of Malacca on their journeys between India and the Far East have noticed that in the Andaman sea, between Nicobar islands and the north east coast of Sumatra, often bands of strongly increased surface roughness (rippings or bands of choppy water) occur [11,34]. Similar observations have been reported in other seas around the globe from time to time. In recent times there has been considerable progress in understanding these kind of observations in terms of internal solitons in the oceans [7]. These studies have been greatly facilitated by photographs taken from satellites and space-crafts orbiting the earth, for example by synthetic aperture radar (SAR) images of ERS-1/2 satellites [34,35].

Peculiar striations of 100 km long, separated by 6 to 15 km and grouped in packets of 4 to 8, visible on satellite photographs (see Fig. 5) of the surface of the Andaman



**Solitons, Tsunamis and Oceanographical Applications of, Figure 5**

SAR image of a 200 km  $\times$  200 km large section of the Andaman Sea acquired by the ERS-2 satellite on April 15, 1996 showing sea surface manifestations of two internal solitary wave packets, see [34]. Figure reproduced from ESA website [www.earth.esa.int/workshops/ers97/papers/alpers3](http://www.earth.esa.int/workshops/ers97/papers/alpers3) with the courtesy of European Space Agency and W. Alpers

and Sulu seas in the Far East, have been interpreted as secondary phenomena accompanying the passage of ‘internal solitons’, which are solitary wavelike distortions of the boundary layer between warm upper layer of sea water and cold lower depths. These internal solitons are traveling edges of warm water, extending hundreds of meters down below the thermal boundary [7]. They carry enormous energy with them which is perhaps the reason for unusually strong underwater currents experienced by deep-sea drilling rigs. Thus these internal solitons are potentially hazardous to sub-sea oil and gas explorations. The ability to predict them can improve substantially the cost effectiveness and safety of offshore drilling.

A systematic study of the underwater currents experienced by an oil rig in the Andaman sea which was drilling at a depth of 3600 ft was carried out by Osborne and Burch in 1980 [11]. They spent four days measuring underwater currents and temperatures. The striations seen on satellite photographs turned out to be kilometer-wide bands of extremely choppy water, stretching from horizon to horizon, followed by about two kilometers of water “as smooth as a millpond”. These bands of agitated water are called “tide rips”, they arose in packets of 4 to 8, spaced about 5 to 10 km apart (they reached the research vessel at approximately hourly intervals) and this pattern was repeated with the regularity of tidal phenomenon.

As described in [7], Osborne and Burch found that the amplitude of each succeeding soliton was less than the previous one, is precisely what is expected for solitons (note that the velocity of a solitary wave solution of KdV equation increases with amplitude, vide Eq. (22)). Thus if a number of solitons are generated together, then we expect them eventually to be arranged in an ordered sequence of decreasing amplitude. From the spacing between successive waves in a packet and the rate of separation calculated from the KdV equation, Osborne and Burch were able to estimate the distance the packet had traveled from its source and thus identify possible source regions [7]. They concluded that the solitons are generated by tidal currents off northern Sumatra or between the islands of the Nicobar chain that extends beyond it and that their observations have good general agreement with the predictions for internal solitons as given by the KdV equation. Numerous recent observations and predictions of solitons in the Andaman sea have clearly established that it is a site where extraordinarily large internal solitons are encountered [34,35].

Further, Apel and Holbrook [36] undertook a detailed study of internal waves in the Sulu sea. Satellite photographs had suggested that the source of these waves was near the southern end of the Sulu sea and their research

ship followed one wave packet for more than 250 miles over a period of two days – an extraordinary coherent phenomenon [7]. These internal solitons travel at speeds of about 8 kilometers per hour (5 miles per hour), with amplitude of about 100 meters and wavelength of about 1700 meters.

Similar observations elsewhere have confirmed the presence of internal solitons in oceans including the strait of Messina, the strait of Gibraltar, off the western side of Baja California, the Gulf of California, the Archipelago of La Maddalena and the Georgia strait [7]. There has also been a number of experimental studies of internal solitons in laboratory tanks in the last few decades [37]. These experiments provide detailed quantitative information usually unavailable in the field conditions, and are also an efficient tool for verifying various theoretical models.

As a theoretical formulation of internal solitons [7], consider two incompressible, immiscible fluids, with densities  $\rho_1$  and  $\rho_2$  and depths  $h_1$  and  $h_2$  respectively such that the total depth  $h = h_1 + h_2$ . Let the lighter fluid of height  $h_1$  be lying over a heavier fluid of height  $h_2$ , in a constant gravitational field (Fig. 6). The lower fluid is assumed to rest on a horizontal impermeable bed, and the upper fluid is bounded by a free surface.

Then as in Sect. “Shallow Water Waves and KdV Type Equations”, we denote the characteristic amplitude of wave by ‘ $a$ ’ and the characteristic wavelength  $l = k^{-1}$ . Then the various nonlinear wave equations to describe the formation of internal solitons follow by suitable modification of the formulation in Sect. “Shallow Water Waves and KdV Type Equations”, assuming viscous effects to be negligible. Each of these equations is completely integrable and admits soliton solutions [7].

(a) **KdV equation (Eq. (21))** follows when

- (i) the waves are of long wavelength  $\delta = \frac{h}{l} \ll 1$ ,
- (ii) the amplitude of the waves are small,  $\epsilon = \frac{a}{h} \ll 1$ , and
- (iii) the two effects are comparable  $\delta^2 = O(\epsilon)$

(b) **Intermediate-Long-Wave (ILW) equation [38]**

$$u_t + u_x + 2uu_x + Tu_{xx} = 0, \quad (39)$$

where  $Tu$  is the singular integral operator

$$(Tf)(x) = \frac{1}{2l} \int_{-\infty}^{\infty} \coth \left\{ \frac{\pi}{2l}(y-x) \right\} f(y) dy \quad (40)$$

with  $\int_{-\infty}^{\infty}$  the Cauchy principal value integral is obtained under the assumption that

- (a) there is a thin (upper) layer,  $\epsilon = \frac{h_1}{h_2} \ll 1$ ,
- (b) the amplitude of the waves is small,  $a \ll h_1$ ,
- (c) the above two effects balance,  $\frac{a}{h_1} = O(\epsilon)$ ,
- (d) the characteristic wavelength is comparable to the total depth of the fluid,  $l = kh = O(1)$  and
- (e) the waves are long waves in comparison with the thin layer,  $kh_1 \ll 1$ .

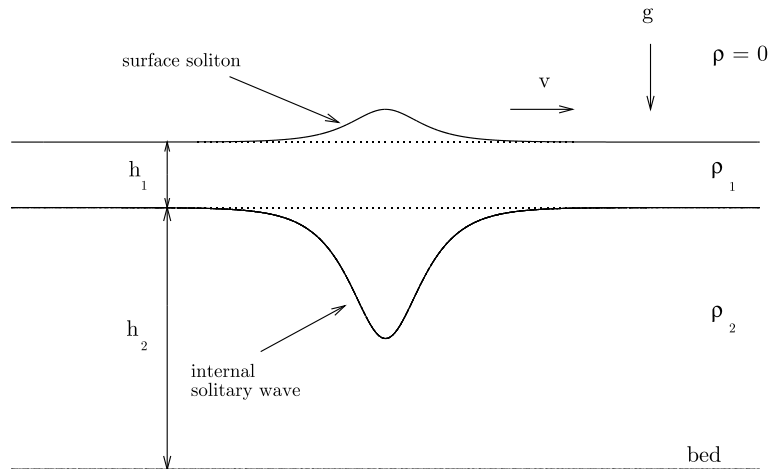
(c) **Benjamin-Ono equation [39]**

$$u_t + 2uu_x + Hu_{xx} = 0, \quad (41)$$

where  $Hu$  is the Hilbert transform

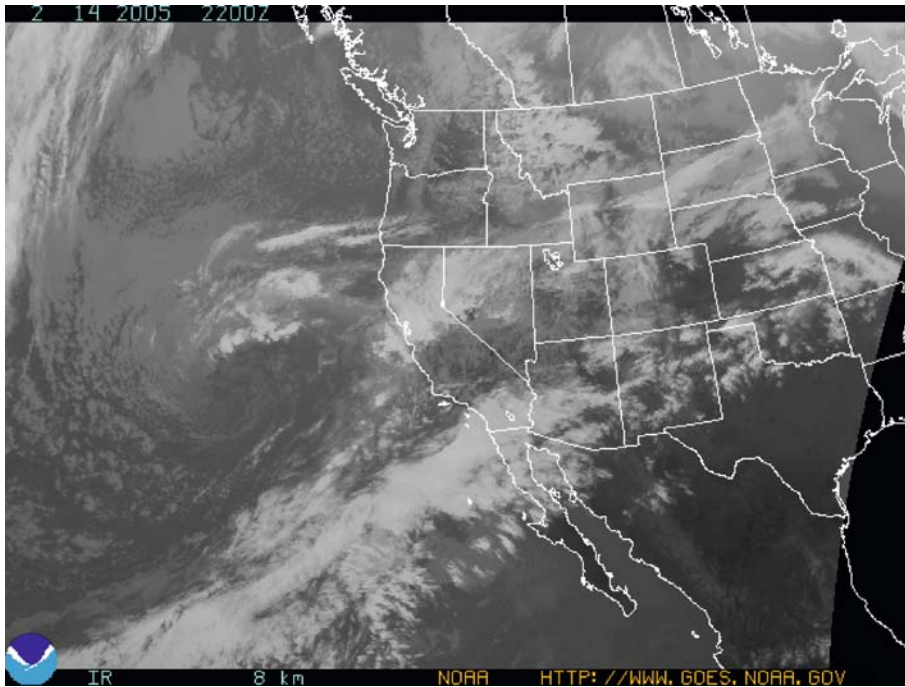
$$(Hf)(x) = \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{f(y)}{y-x} dy, \quad (42)$$

is obtained under the assumption



Solitons, Tsunamis and Oceanographical Applications of, Figure 6

Formation of internal soliton (note that under suitable conditions small amplitude surface soliton can also be formed)



#### Solitons, Tsunamis and Oceanographical Applications of, Figure 7

A coupled high/low pressure systems in the form of a Rossby soliton formed off the coast of California/Washington on Valentine's Day 2005. The low pressure front hovers over Los Angeles dumping 30 inches of rain on the city in two weeks. The high pressure system lingers off the coast of Washington state providing unseasonably warm and sunny weather. This particular Rossby soliton proved exceptionally stable because of its remarkable symmetry. In an accompanying animated gif in this website one can watch the very interesting phenomenon as the jet stream splits around the soliton suctioning warm wet air from the off the coast of Mexico to Arizona, leaving behind a welcome drenching of rain. The figure and caption have been adapted from the website [http://mathpost.la.asu.edu/~rubio/rossby\\_soliton/rs.html](http://mathpost.la.asu.edu/~rubio/rossby_soliton/rs.html) with the courtesy of the National Oceanic and Atmospheric Administration (NOAA) and Antonio Rubio

- (a) there is a thin (upper) layer  $h_1 \ll h_2$ ,
- (b) the waves are long waves in comparison with the thin layer,  $kh_1 \ll 1$ ,
- (c) the waves are short in comparison with the total depth of the fluid,  $kh \gg 1$  and
- (d) the amplitude of the waves is small,  $a \ll h_1$ .

It may be noted that in the shallow water limit, as  $\delta \rightarrow 0$ , the ILW equation reduces to the KdV equation, while the Benjamin-Ono equation reduces to it in the deep water wave limit as  $\delta \rightarrow \infty$ . Each of these equations have their own ranges of validity and admit solitary wave and soliton solutions to represent internal solitons of the oceans.

#### Rossby Solitons

The atmospheric dynamics is an important subject of human concern as we live within the atmosphere and are continuously affected by the weather and its rather complex behavior. The motion of the atmosphere is intimately

connected with that of the ocean with which it exchanges fluxes of momentum, heat and moisture [40]. Then the atmospheric dynamics is dominated by the rotation of the earth and the vertical density stratification of the surrounding medium, leading to newer effects. Similar effects are also present in other planetary dynamics as well.

In the atmosphere of a rotating planet, a fluid particle is endowed with a certain rotation rate (Coriolis frequency), determined by its latitude. Consequently its motion in the north-south direction is inhibited by conservation of angular momentum. The large scale atmospheric waves caused by the variation of the Coriolis frequency with latitude are called *Rossby waves*. In Sect. "Internal Solitons" we saw that KdV equation and its modifications model internal waves. Since there is a resemblance between internal waves and Rossby waves, it is expected that KdV like equations can model Rossby waves as well [8]. Under the simplest assumptions like long waves, incompressible fluid,  $\beta$ -plane approximations, etc. Benney [41] had derived the KdV equation as a model for Rossby

waves in the presence of east-west zonal flow. Boyd [42] had shown that long, weakly nonlinear, equatorial Rossby waves are governed either by KdV or MKdV equation. Recently it has been shown by using the eight years of Topex/Poseidon altimeter observations [43] that a detailed characterization of major components of Pacific dynamics confirms the presence of equatorial Rossby solitons.

Also observational and numerical studies of propagation of nonlinear Rossby wave packets in the barotropic atmosphere by Lee and Held [44,45] have established their presence, notably in the Northern Hemisphere as storm tracks, but more clearly in the Southern Hemisphere (see Fig. 7). They also found that the wavepackets both in the real atmosphere and in the numerical models behave like the envelope solitons of the nonlinear Schrödinger equation (Fig. 7).

An interesting application of KdV equation to describe Rossby waves is the conjecture of Maxworthy and Redekopp [46] that the planet Jupiter's Great Red Spot might be a solitary Rossby wave. Photographs taken by the spacecraft Voyager of the cloud pattern show that the atmospheric motion on Jupiter is dominated by a number of east-west zonal currents, corresponding to the jet streams of the earth's atmosphere [8]. Several oval-shaped spots are also seen, including the prominent Great Red Spot in the southern hemisphere. The latter one has been seen approximately this latitude for hundreds of years and maintains its form despite interactions with other atmospheric objects. One possible explanation is that the Red Spot is a solitary wave/soliton of the KdV equation, deduced from the quasigeostrophic form of potential vorticity equation for an incompressible fluid. A second test of the model is the combined effect of interaction of the Red Spot and Hollow of the Jupiter atmosphere on the South Tropical Disturbance which occurred early in the 20th century and lasted several decades. Maxworthy and Redekopp [46] have interpreted this interaction as that of two soliton collision of KdV equation, with the required phase shift [6,7,8].

The above connection between the KdV solitary wave and the rotating planet may be seen more concretely by the following argument as detailed in [8]. One can start with the quasigeostrophic form of the potential vorticity equation for an incompressible fluid in the form

$$\left\{ \left( \frac{\partial}{\partial t} + v \frac{\partial}{\partial x} \right) + \epsilon \left( \frac{\partial \psi}{\partial y} \right) \left( \frac{\partial}{\partial x} - \frac{\partial \psi}{\partial x} \frac{\partial}{\partial y} \right) \right\} \left\{ \mu^2 \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial}{\partial z} \left( K^2 \frac{\partial}{\partial z} \right) \right\} \psi + (\beta - U'') \frac{\partial \psi}{\partial x} = 0, \quad (43)$$

where  $x$ ,  $y$  and  $z$  represent the east, north and vertical directions, while the function  $U$  is related to the horizontal stream function  $\Phi$  as  $\Phi(x, y, z, t) = \int_y^y U(\eta) d\eta + \epsilon \psi(x, y, z, t)$  in terms of a zonal shear flow and a perturbation. In (43), in the  $\beta$ -plane approximation, the Coriolis parameter is given as  $f = 2\Omega \sin \theta_0 + \beta y$  and the function  $K(z)$  is given by  $K(z) = 2\Omega \sin \theta_0 l_2 / N(z) d$ , where  $N(z)$  is the Brunt-Väisälä frequency and  $l_2$  is the characteristic length scales in the north-south direction while  $d$  is the length scale in the vertical direction. Note that  $K$  compares the effects of rotation and density of variation. Then  $\mu = l_2/l_1$  represents the ratio of the length scales in the north-south and east-west directions.

In the linear ( $\epsilon \rightarrow 0$ ), long wave ( $\mu \rightarrow 0$ ) limit, the potential function  $\psi$  has the form  $\psi = \sum_n A_n(x - c_n t) \phi_n(y) p_n(z)$ , where  $c_n$  is deduced by solving two related eigenvalue problems [8],  $(K^2 p_n')' + k_n^2 p_n = 0$ ,  $p_n(0) = p_n(1) = 0$  and  $\phi_n'' - k_n^2 \phi_n + [(\beta - U'')/(U - c_n)] \phi_n = 0$ ,  $\phi_n(y_s) = \phi_n(y_N) = 0$ . If the various modes are separable on a short time scale, one can then deduce an evolution equation for the individual modes, by eliminating secular terms at higher order in the expansion. Depending on the nature and existence of a stable density of stratification, which is characterized by the function  $N(z)$  mentioned above, either KdV or mKdV equation can be deduced for a given mode and thereby establishing the soliton connection in the present problem.

In spite of the complexity of the phenomenon underlying Rossby solitons there is clear evidence of the significance of solitonic picture.

## Bore Solitons

Rivers which flow into the open oceans are usually affected by tidal flows, tsunami or storm surge. For a typical estuary as one moves towards the mouth of the river, the depth increases and width decreases. When a tidal wave, tsunami or storm surge hits such an estuary, it can be seen as a hydraulic jump (step-wise perturbations like a shock-wave) in the water height and speed which will propagate upstream [47]. Far less dangerous but very similar is the bore (mascaret in French), a tidal wave which can propagate in a river for considerable distances.

Typical examples of bores occur in Seine river in France and the Hooghli river in West Bengal in India. A bore existed in the Seine river upto 1960 and disappeared when the river was dredged. The tide amplitude here is one of the largest in the world. The Hughli river is a branch of the Ganges that flows through Kolkata where a bore of 1m is present, essentially due to the shallowness of the river. Another interesting example is that at the time





**Solitons, Tsunamis and Oceanographical Applications of, Figure 8**

**Tidal bore at the mouth of the Araguari River in Brazil. The bore is undular with over 20 waves visible. (Adapted from the book "Gravity Currents in the Environment and the Laboratory" by John E. Simpson, Cambridge University Press with the courtesy of D.K. Lynch, John E. Simpson and Cambridge University Press)**

of the 1983 Japan sea tsunami, waves in the form of a bore ascended many rivers [48]. In some cases, bores had the form of one initial wave with a train of smaller waves and in other cases only a step with flat water surface behind was observed (Fig. 8). The Hangzhou bore in China is a tourist attraction. Other well known bores occur in the Amazon in Brazil and in Australia.

Another interesting situation where bore solitons were observed was in the International H<sub>2</sub>O Project (IHOP), as a density current (such as cold air from thunderstorm) intrudes into a fluid of lesser density that occurs beneath a low level inversion in the atmosphere. A spectacular bore and its evolution into a beautiful amplitude-ordered train of solitary waves were observed and sampled during the early morning of 20 June 2002 by the Leandre-II abroad the P-3 aircraft. The origin of this bore was traceable to a propagating cold outflow boundary from a mesoscale convective system in extreme western Kansas [49].

In the process of propagation the bore undergoes dissipation, dispersive disintegration, enhancement due to decrease of the river width and depth, influence of nonlinear effects and so on. The profile depends on the Froude number which is a dimensionless ratio of inertial and gravitational effects. Theoretical models have been developed to study these effects based on KdV and its generalizations [14]. For example, bore disintegration into solitons in channel of constant parameters can be studied in signal coordinates in terms of the KdV like equation for the perturbation of water surface,

$$\eta_x + \frac{1}{c_0}(1 - \alpha\eta)\eta_t - \beta\eta_{ttt} = 0, \quad (44)$$

where  $c_0 = \sqrt{gh}$ ,  $\alpha = 3/2h$ ,  $\beta = h^2/6c_0^3$ ,  $h$  being the depth of the river, with the bore represented by a Heaviside

step function as the boundary condition at  $x = 0$ . Other effects then can be incorporated into a variable KdV equation of the form (26).

### Future Directions

We have indicated a few of the most important oceanographical applications of solitons including tsunamis, internal solitons, Rossby solitons and bore solitons. There are other interesting phenomena like capillary wave solitons [15], resonant three and four wave interaction solitons [16], etc. which are also of considerable interest depending on whether the wave propagation corresponds to shallow, intermediate or deep waters. Whatever be the situation, it is clear that experimental observations as well as their theoretical formulation and understandings are highly challenging complex nonlinear evolutionary problems. The main reason is that the phenomena are essentially large scale events and detailed experimental observations require considerable planning, funding, technology and manpower. Often satellite remote sensing measurements need to be carried out as in the case of internal solitons and Rossby solitons. Events like tsunami propagation are rare and time available for making careful measurements are limited and heavily dependent on satellite imaging, warning systems and after event measurements. Consequently developing and testing theoretical models are extremely hazardous and difficult. Yet the basic modeling in terms of solitary wave/soliton possessing nonlinear dispersive wave equations such as the KdV and NLS family of equations and their generalizations present fascinating possibilities to understand these large scale complex phenomena and opens up possibilities of prediction. Further understanding of such nonlinear evolution equations, both integrable and nonintegrable systems particularly in higher dimensions, can help to understand the various phenomena clearly and provide means of predicting events like tsunamis and damages which occur due to internal solitons and bores. Detailed experimental observations can also help in this regard. It is clear that what has been understood so far is only the qualitative aspects of these phenomena and much more intensive work is needed to understand the quantitative aspects to predict them.

### Bibliography

#### Primary Literature

1. Russel JS (1844) Reports on Waves. 14th meeting of the British Association for Advancement of Science. John Murray, London, pp 311–390



2. Bullough RK (1988) The Wave Par Excellence. The solitary progressive great wave of equilibrium of fluid: An early history of the solitary wave. In: Lakshmanan M (ed) *Solitons: Introduction and Applications*. Springer, Berlin
3. Korteweg DJ, de Vries G (1895) On the change of form of long waves advancing in a rectangular canal, and on a new type of long stationary waves. *Philos Mag* 39:422–443
4. Zabusky NJ, Kruskal MD (1965) Interactions of solitons in a collisionless plasma and recurrence of initial states. *Phys Rev Lett* 15:240–243
5. Gardner CS, Greene JM, Kruskal MD, Miura RM (1967) Method for solving the Korteweg–de Vries equation. *Phys Rev Lett* 19:1095–97
6. Lakshmanan M, Rajasekar S (2003) *Nonlinear Dynamics: Integrability and Chaos*. Springer, Berlin
7. Ablowitz MJ, Clarkson PA (1991) *Solitons, Nonlinear Evolution Equations and Inverse Scattering*. Cambridge University Press, Cambridge
8. Ablowitz MJ, Segur H (1981) *Solitons and Inverse Scattering Transform*. Society for Industrial and Applied Mathematics, Philadelphia
9. Scott AC (1999) *Nonlinear Science: Emergence and Dynamics of Coherent Structures*. Oxford University Press, New York
10. Dudley WC, Miu L (1988) *Tsunami!* University of Hawaii Press, Honolulu
11. Osborne AR, Burch TL (1980) Internal solitons in the Andaman Sea. *Science* 258:451–460
12. Rossby GG (1939) Relation between variations in the intensity of the zonal circulation of the atmosphere. *J Mar Res* 2:38–55
13. Redekopp L (1977) On the theory of solitary Rossby waves. *J Fluid Mech* 82:725–745
14. Caputo JG, Stepanyants YA (2003) Bore formation and disintegration into solitons in shallow inhomogeneous channels. *Nonlinear Process Geophys* 10:407–424
15. Longuet-Higgins MS (1993) Capillary gravity waves of solitary type and envelope solitons in deep water. *J Fluid Mech* 252:703–711
16. Philips OM (1974) Nonlinear dispersive waves. *Ann Rev Fluid Mech* 6:93–110
17. Helal MA, Molines JM (1981) Nonlinear internal waves in shallow water. A theoretical and experimental study. *Tellus* 33:488–504
18. Benjamin TB, Bona JL, Mahoney JJ (1972) Model equations for long waves in nonlinear dispersive systems. *Philos Trans A Royal Soc* 272:47–78
19. Camassa R, Holm D (1992) An integrable shallow water equation with peaked solitons. *Phys Rev Lett* 71:1661–64
20. Zakharov VE (1968) Stability of periodic waves of finite amplitude on the surface of a deep fluid. *J Appl Mech Tech Phys* 2:190–194
21. Hasimoto H, Ono H (1972) Nonlinear modulation of gravity waves. *J Phys Soc Jpn* 33:805–811
22. Benney DJ, Roskes G (1969) Wave instabilities. *Stud Appl Math* 48:377–385
23. Davey A, Stewartson K (1974) On three dimensional packets of surface waves. *Proc Royal Soc Lond A* 338:101–110
24. Fokas AS, Santini PM (1990) Dromions and a boundary value problem for the Davey–Stewartson I equation. *Physica D* 44:99–130
25. Kundu A (ed) (2007) *Tsunami and Nonlinear Waves*. Springer, Berlin
26. Dias F, Dutykh D (2007) Dynamics of tsunami waves. In: Ibrahimbegovic A, Kozar I (eds) *Extreme man-made and natural hazards in dynamics of structures*. NATO security through Science Series. Springer, Berlin, pp 201–224
27. Dutykh D, Dias F (2007) Water waves generated by a moving bottom. In: Kundu A (ed) *Tsunami and Nonlinear Waves*. Springer, Berlin, pp 65–94
28. Okada Y (1992) Internal deformation due to shear and tensile faults in a half space. *Bull Seism Soc Am* 82:1018–1040
29. Carrier GF, Wu TT, Yeh H (2003) Tsunami runup and drawdown on a plane beach. *J Fluid Mech* 475:79–99
30. Banerjee P, Politz FF, Burgman R (2005) The size and duration of the Sumatra–Andaman earthquake from far-field static offsets. *Science* 308:1769–1772
31. Constantin A, Johnson RS (2006) Modelling tsunamis. *J Phys A* 39:L215–L217
32. Segur H (2007) Waves in shallow water, with emphasis on the tsunami of (2004). In: Kundu A (ed) *Tsunami and Nonlinear Waves*. Springer, Berlin, pp 3–29
33. Lakshmanan M (2007) Integrable nonlinear wave equations and possible connections to tsunami dynamics. In: Kundu A (ed) *Tsunami and Nonlinear Waves*. Springer, Berlin, pp 31–49
34. Alpers W, Wang-Chen H, Cook L (1997) Observation of internal waves in the Andaman Sea by ERS SAR. *IEEE Int J* 4:1518–1520
35. Hyder P, Jeans DRG, Cauquill E, Nerzic R (2005) Observations and predictability of internal solitons in the northern Andaman Sea. *Appl Ocean Res* 27:1–11
36. Apel JR, Holbrook JR (1980) The Sulu sea internal soliton experiment, 1. Background and overview. *EOS Trans AGU* 61:1009
37. Ostrovsky LA, Stepanyants YA (2005) Internal solitons in laboratory experiments: Comparison with theoretical models. *Chaos* 15(1–28):037111
38. Joseph RI (1977) Solitary waves in a finite depth fluid. *J Phys A* 10:L225–L227
39. Benjamin TB (1967) Internal waves of permanent form in fluids of great depth. *J Fluid Mech* 29:559–592
40. Kundu PK, Cohen IM (2002) *Fluid Mechanics*, Second Edition. Academic Press, San Diego
41. Benney DJ (1966) Long nonlinear waves in fluid flows. *Stud Appl Math* 45:52–63
42. Boyd JP (1980) Equatorial solitary waves. Part I: Rossby solitons. *J Phys Oceanogr* 10:1699–1717
43. Susanto RD, Zheng Q, Xiao-Hai Y (1998) Complex singular value decomposition analysis of equatorial waves in the Pacific observed by TOPEX/Poseidon altimeter. *J Atmospheric Ocean Technol* 15:764–774
44. Lee S, Held I (1993) Baroclinic wave packets in models and observations. *J Atmospheric Sci* 50:1413–1428
45. Tan B (1996) Collision interactions of envelope Rossby solitons in a barotropic atmosphere. *J Atmospheric Sci* 53:1604–1616
46. Maxworthy T, Redekopp LG (1976) Theory of the Great Red Spot and other observed features of the Jovian atmosphere. *Icarus* 29:261–271
47. Caputo JG, Stepanyants YA (2007) Tsunami surge in a river: a hydraulic jump in an inhomogeneous channel. In: Kundu A (ed) *Tsunami and Nonlinear Waves*. Springer, Berlin, pp 97–112
48. Tsuji T, Yanuma T, Murata I, Fujiwara C (1991) Tsunami ascending in rivers as an undular bore. *Nat Hazard* 4:257–266
49. Koch SE, Pagowski M, Wilson JW, Fabry F, Flamant C, Feltz W, Schwemmer G, Geerts B (2005) The structure and dynamics of

atmospheric bores and solitons as determined from remote sensing and modelling experiments during IHOP, AMS 32nd Conference on Radar Meteorology, Report JP6J.4

## Books and Reviews

- Lamb H (1932) *Hydrodynamics*. Dover, New York
- Miles JW (1980) Solitary waves. *Ann Rev Fluid Mech* 12:11–43
- Stoker JJ (1957) *Water Waves*. Interscience, New York
- Dauxois T, Peyrard M (2006) *Physics of Solitons*. Cambridge University Press, Cambridge
- Johnson RS (1997) *An Introduction to the Mathematical Theory of Water Waves*. Cambridge University Press, Cambridge
- Hammack JL (1973) A note on tsunamis: their generation and propagation in an ocean of uniform depth. *J Fluid Mech* 60:769–800
- Drazin PG, Johnson RS (1989) *Solitons: An Introduction*. Cambridge University Press, Cambridge
- Mei CC (1983) *The Applied Dynamics of Ocean Surface Waves*. Wiley, New York
- Scott AC et al (1973) The soliton: a new concept in applied science. *Proc IEEE* 61:1443–83
- Scott AC (ed) (2005) *Encyclopedia of Nonlinear Science*. Routledge, New York
- Sulem C, Sulem P (1999) *The Nonlinear Schrödinger Equation*. Springer, Berlin
- Helfrich KR, Melville WK (2006) Long nonlinear internal waves. *Ann Rev Fluid Mech* 38:395–425
- Bourgault D, Richards C (2007) A laboratory experiment on internal solitary waves. *Am J Phys* 75:666–670

## Space Plasmas, Dynamical Complexity in

TOM CHANG

Kavli Institute for Astrophysics and Space Research,  
Massachusetts Institute of Technology, Cambridge, USA

## Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Plasma Resonances and Coherent Structures  
 Coarse-Grained Dissipation  
 and Magnetic Reconfiguration  
 Non-Gaussian Probability Distribution Functions  
 Wavelet Transforms and Intermittency  
 Multifractals  
 Invariant Scaling  
 Dynamical Modeling – The Lu–Klimas Magnetic Field  
 Reversal Model  
 Future Directions  
 Acknowledgments  
 Bibliography

## Glossary

- Alfvén wave** Transverse magnetohydrodynamic wave that propagates along the magnetic field line direction in an electrically conducting fluid.
- Anomalous resistivity** Nonclassical resistivity in plasma dynamics.
- Astronomical unit (AU)** Unit of length nearly equal to the semi-major axis of Earth's orbit around the Sun.
- Auroral electrojet index (AE)** Empirical index that measures magnetic perturbations in the auroral zone. It provides a gauge for the horizontal current strength in the auroral ionosphere.
- Auroral zone** Region where auroral activities are observed.
- Castaing distribution** A convolution of Gaussians with variances distributed according to a log-normal distribution.
- Chaos** Nonsteady, nonperiodic, complex dynamical motion.
- Coarse-grained dissipation** Apparent dissipation due to interactions of coherent structures.
- Coarse-graining** Replacement of a detailed description with a lower-resolution coarse-grained model.
- Coherent structure** Large scale structure of nonlinearly interacting dynamical systems.
- Correlation function** Ensemble average of the correlation of fluctuations.
- Criticality** Behavior of extended system at which scale-invariance prevails.
- Crossover** Transition from one critical state to another or multitudes of critical states; or, transition from one fractal state to another or multitudes of fractal states.
- Current sheet** Sheet-like region with strong current density usually generated by strong local magnetic shear.
- Cusp** Funnel shaped region dividing the sunward and tailward sides of the magnetic field lines of the Earth.
- Drift Alfvén vortex** Field-aligned coherent Alfvénic vortex structure that drifts across the magnetic field.
- Dynamical complexity** Complex stochastic behavior of numerous nonlinearly interacting coherent structures of various sizes.
- Dynamical renormalization group** Group generated by coarse-graining scale transformations of a nonlinear dynamical system.
- Extended similarity scaling (ESS)** Scale invariant property among the structure or partition functions of a fluctuating event.
- Exponent relation** Algebraic relation among scaling exponents.
- Extreme value theory** Statistical theory dealing with extreme deviations from the mean.

- Flatness (kurtosis)** A normalized scale dependent measure related to the fourth order moment of the scale dependent coarse-grained probability distribution. It may be evaluated by taking the ensemble average of the square of the normalized power using the wavelet transform.
- Forced and/or self-organized criticality (FSOC)**  
Organization of a dynamical system into a scale invariant critical state with or without some tuning or forcing.
- Fractal** Fluctuations that exhibit generalized dimension of an irrational number.
- Gaussian distribution** Normal distribution.
- Generalized dimension** Dimension of the measure of fractal characteristics of a fluctuating event based on the limiting characteristic of the logarithm of the partition function at small scales.
- Helicity** Extent to which the magnetic field wraps around itself.
- Heliosphere** Magnetic bubble containing our solar system, including the solar wind.
- Hysteresis** History dependence of physical system that dissipates energy.
- Intermittency** Random fluctuations that exhibit non-self-similar characteristics.
- Irrelevant parameter** Parameter that becomes irrelevant under repeated coarse-graining transformations.
- Ionosphere** Upper most portion of atmosphere that is ionized and connected to the magnetosphere.
- Local intermittency measure, Lim** Normalized power in wavelet transform that is intensity, scale and location dependent.
- Low-dimensional chaos** Chaotic motion characterized by a small number of parameters.
- Lu/Klimas model** One-dimensional dynamical model with anomalous diffusion involving hysteresis, which exhibits scale invariant stochastic processes.
- Magnetic reconfiguration (and/or reconnection)**  
Reconfiguration of magnetic topology, which may or may not involve the phenomenon of magnetic reconnection of field lines.
- Magnetohydrodynamics** Mathematical description of electrically conducting fluids moving within electromagnetic environments.
- Magnetosphere** Region of space around the Earth (or any astrophysical object), that is controlled by its magnetic field.
- Magnetotail** Tail region (away from the Sun) of the magnetosphere of the Earth or any object in the heliosphere.
- Multifractals** Fluctuations characterized by multitudes of fractal dimensions that are not linearly related.
- P-model** Two-scale cascade model with “p” as the parameter that characterizes the fragmentation probability of the cascading process of intermittency.
- Partition function of order  $q$**   $q$ th moment-order of the segmental coarse-grained probabilities.
- PDF** Probability distribution function.
- Plasma resonance** Site at which a particular plasma propagation mode vanishes.
- Rank-ordered multifractal analysis, ROMA** Analysis of multifractal characteristics based on the ordering of fractal dimensions in terms of scaled sizes of the intermittent fluctuations.
- Relevant parameter** Parameter that becomes more and more relevant under repeated coarse-graining transformations.
- Response function** Ensemble average of the response of the system due to fluctuations.
- Scale invariants** Quantities that do not vary under coarse-graining scale transformations.
- Singularity spectrum  $f(\alpha)$**  Continuum distribution of differential measures with singularity index  $\alpha$ .
- Self-organized criticality (SOC)** Auto-organization of a dynamical system into a scale-invariant critical state without significant tuning.
- Solar wind** Stream of charged particles in interplanetary space ejected from the upper atmosphere of the Sun.
- Space plasma** Ionized medium in the space environment.
- Structure function of order  $q$**   $q$ th order moment of the probability distribution function of scale-dependent, coarse-grained fluctuations. It may be evaluated in terms of the ensemble average of the  $q$ th power of the scale-dependent fluctuations.
- Taylor hypothesis** Hypothesis that interprets temporal fluctuations as spatial fluctuations when the transit time of the fluctuations is much less than the characteristic evolution time. Also called Taylor’s “frozen-in” hypothesis.
- Turbulence** Chaotic stochastic flow of a continuum or discrete medium with infinite number of degrees of freedom.
- Wavelet transform** Transform generated by a basis set of wave-packet-like functions, which provides information of intensity, location and scale of a fluctuating quantity.
- Whistler wave** Right-hand polarized electromagnetic wave in a plasma with frequency well below the electron cyclotron frequency and above the ion cyclotron frequency. The high frequency whistler waves travel

faster than the low frequency ones along the magnetic field lines in the plasma medium.

## Definition of the Subject

Commonly called the fourth state of matter, plasmas make up 99 per cent of the Universe. They are the major constituents of the solar-terrestrial, interstellar, and intergalactic media – generally in the form of ionized gases. According to Tonks [118] who co-authored the celebrated Tonks and Langmuir [119] paper on oscillations in ionized gases, Langmuir [70] coined the terminology “plasma” to describe, in today’s language, a quasineutral mixture – probably because of the likeness between the electrons and ions in an ionized medium and the red and white corpuscles in a blood plasma [99]. Recent observations, particularly in-situ measurements in the heliosphere and solar-terrestrial regions, have indicated that space plasmas commonly exhibit random, intermittent, and anisotropic fluctuating characteristics at all spatiotemporal scales while interacting with the electromagnetic environment. Such observations have led theoreticians and experimentalists alike to regard the stochastic behaviors of space plasmas as prime candidates of naturally occurring examples of dynamical complexity.

By definition, “dynamical complexity” is a phenomenon exhibited by a nonlinearly interacting dynamical system within which multitudes of different sizes of large scale “coherent structures” are formed, resulting in a complicated global nonlinear stochastic behavior for the dynamical system – vastly different from that could be surmised from the original dynamical equations.

The main purpose of this article is to demonstrate the prevalence of dynamical complexity in space plasmas and to indicate the various statistical methods of analyzing this kind of dynamical processes as well as to discuss theoretical and numerical methods that can provide the basic understanding of such processes.

## Introduction

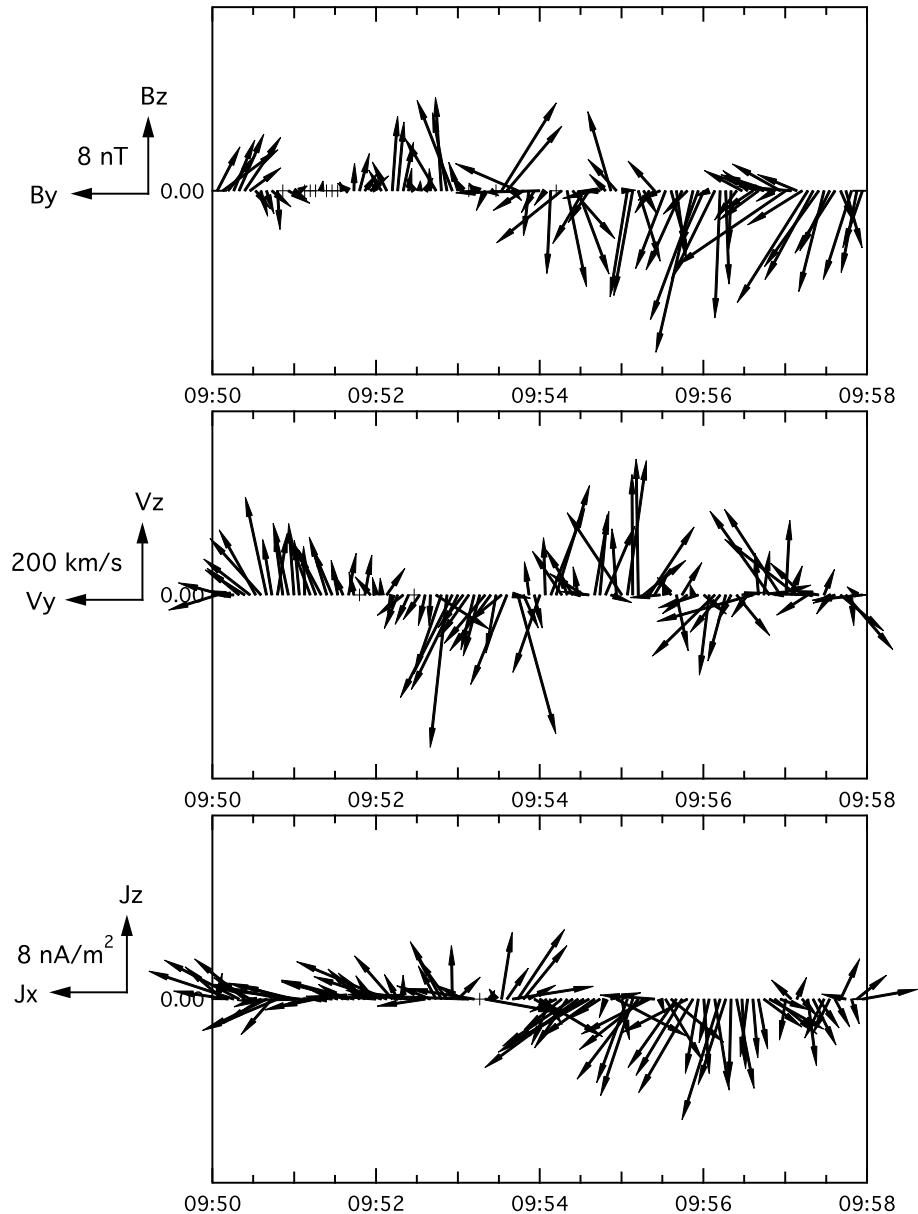
Traditionally, analyzes of space plasma processes are based on fluid or kinetic formulations. Among the fluid descriptions, the simplest is that based on the so-called one-component magnetohydrodynamic (MHD) equations, which includes the equations of motion, continuity, and energy conservation, the Ohm’s law and the Maxwell equations for a single conducting fluid medium (see, e. g. [7]). Associated with such a formulation are the concepts of magnetic field lines, streamlines and other continuum variables such as velocity, electric and magnetic fields, as well as plasma, charge and current densities. The dynamical

state of the system is understood in terms of a topology generally characterized by the relatively smooth or piecewise continuous variations of such entities with space and time.

Most of the observed space plasma processes, however, generally exhibit discernible turbulent fluctuations of such quantities. The standard approach to the theoretical analyzes of dynamically fluctuating states is based on the concepts of linear instabilities, nonlinear growths and interactions of the wave modes. Although the basic governing equations generally contain strong nonlinearities, one is led to believe that turbulent motions may be understood by expressing the fluctuations in Fourier modes (plane waves) and then considering the interactions among these nonlocal modes by requiring them to satisfy the “basic” equations. This procedure produced such intractable complications that have led to decades of futile search for a workable theory of “turbulence”.

In reality, turbulent fluctuations in space plasmas are generally composed of the simultaneous coexistence of propagating modes and intermittent nonlinearly interacting spatiotemporal structures [32,33]. The “physics” of the bimodal state of such type of admixture of turbulent fluctuations may be understood from the point of view of the development and interactions of coherent structures arising from plasma resonance sites and the ensuring dynamical complexity resulting from such developments and interactions. In this article, we shall consider the dynamical complexity in space plasmas from such a concept. Sample results of direct numerical simulations and dynamical modeling including the calculated fluctuation probability distribution functions and local intermittency measures based on the wavelet transforms are provided to characterize the sporadic, localized, and scale-dependent nature of the intermittent turbulence. The concepts of multifractals, scale invariance and symmetry-breaking will be introduced. Observational examples demonstrating the existence of such phenomena and the associated invariance properties of intermittent space plasma turbulence will be provided.

Applications of the ideas of dynamical complexity to space plasmas were relatively recent. The first contact of these ideas with space plasmas was probably contained in a series of papers addressing the apparent low-dimensional chaotic behavior related to the dynamics of the Earth’s magnetosphere [4,65,101,103,104,105]. These ideas were followed by a paper by Chang [22] that suggested the possibility of interpreting such phenomenon from the point of view of forced and/or self-organized criticality (FSOC) surmised from the method of the dynamic renormalization group (DRG). This suggestion was moti-



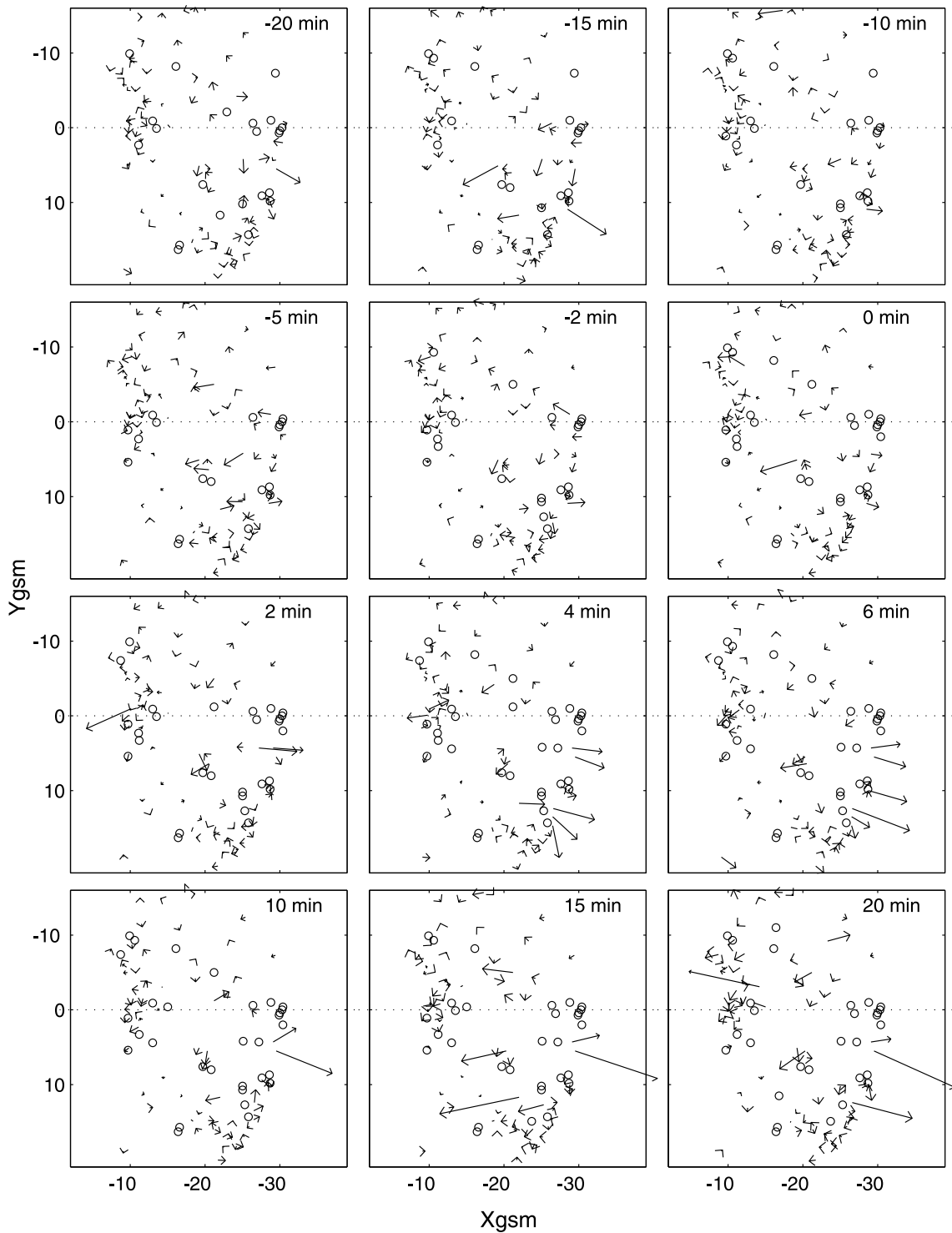
Space Plasmas, Dynamical Complexity in, Figure 1

Vector representation of the components of the magnetic field ( $B_y$ ,  $B_z$ ), ion plasma flow ( $V_y$ ,  $V_z$ ), and current density ( $J_x$ ,  $J_z$ ) during a magnetic substorm (i. e., when there was intense auroral activity) on August 22, 2001 observed by the Cluster spacecraft at a downstream distance of about 19 Re (Earth radii) in the magnetotail showing the large variability of these plasma parameters. Reprinted from [82] with thanks to the European Geophysical Union

vated by the low-dimensional chaos papers as well as the interpretation of the observed large variability of plasma fluctuations in the Earth's magnetotail as the result of sporadic and localized current disruptions [78], Figs. 1 and 2. At approximately the same time, Lu and Hamilton [74] used the idea of self-organized criticality (SOC) of Bak et al. [3] to explain the curious scale-independent behavior

of the solar flare occurrence rate on the flare sizes. Such ideas suggest that complex dynamical systems generally organize themselves into states with statistical properties describable by power laws. (A succinct tutorial review of the SOC concept may be found in Jensen [63]). An alternative viewpoint of dynamical complexity is based on the concept of multifractals [54,98] which is a generaliza-





#### Space Plasmas, Dynamical Complexity in, Figure 2

A temporal sequence of synoptic patterns of total plasma flow showing large variabilities constructed from superposed epoch analysis of 102 magnetic substorm events from Geotail spacecraft observations. The scale is 100 km/s per Re (Earth radius). The acronym GSM stands for Geocentric Solar Magnetospheric coordinate system. The X-axis points to the Sun. The XZ-plane contains the Earth's dipole axis, and the Y-axis completes the right-handed coordinate system in units of Re. Reprinted from [80] with thanks to the American Geophysical Union

tion of the original ideas of fractals introduced by Mandelbrot [83]. Burlaga [12,13,14] was the first to incorporate such ideas to interpret the intermittent turbulent behavior of the solar wind. Other solar wind studies based on such a point of view followed, e.g. [16,17,18,19,20,58,59,60,84,102]. These ideas were followed by Consolini et al. [41] and Consolini [37] in their interpretations of the observed fractal time series of the Auroral Electrojet (AE) indices.

Since then, a flurry of activities has blossomed into the current investigations of the phenomenon of complexity in space plasmas. Perhaps the most noteworthy are the connection of such phenomena with the traditional ideas of intermittent turbulence in plasmas and the development of the concept of crossovers (transitions) from one criticality to another and onto multifractals [27,28]. This article will provide simple descriptions of some of these modern ideas in conjunction with actual observational results and numerical simulations.

Due to the limitation of space, this entire article will center on the discussion of the complexity of space plasmas from the fluid description. It should be understood that situations also exist in space plasmas where the physical phenomena are related to the complexity of kinetic plasma effects, see, e.g., [33,111,115]. Since the purpose of this article is to introduce to the readers the basic concepts of – and recent developments in – the physics of dynamical complexity in space plasmas, only those primary references that touch upon these introductory ideas are included. No attempt has been made to include the hundreds of outstanding contributions that are available on the subject.

The contents of this article are structured as follows. In the next section, the concepts of plasma resonances and coherent structures are introduced. This is followed by a discourse on coarse-grained dissipation and magnetic reconfiguration. The ideas of non-Gaussian probability distributions of turbulent fluctuations, wavelet transforms and intermittency are then discussed in the two subsequent sections. Next, we introduce the basic ideas of multifractals in terms of structure functions, partition functions, singularity measures and rank-ordered multifractal spectra in Sect. “[Multifractals](#)”. The concepts of scale invariants, forced and/or self-organized criticality are described in Sect. “[Invariant Scaling](#)” along with a discourse relating these ideas to multifractals. In Sect. “[Dynamical Modeling – The Lu-Klimas Magnetic Field Reversal Model](#)”, we discuss the utility of dynamical models. A simple one-dimensional model, the Lu-Klimas model, is provided to illustrate the usefulness of such methods. This is followed by a brief discussion of what future en-

tails in the study and research of dynamical phenomena of complexity related to space plasmas. In addition to the primary list of references that are cited in the article, we also provide a secondary list of review articles and books in the bibliography for further in-depth study.

## Plasma Resonances and Coherent Structures

Plasmas are known for their ability to form numerous types of coherent structures of varied sizes with scales that are generally much larger than those of the constituent particles (ions, electrons, neutrals) of the plasma medium. The reason behind this is closely related to the nonlinear and long range forces of the dynamical couplings among the constituent particles and species interacting within the electromagnetic environment.

Examples of coherent structures abound in the literature on theories and observations of nonlinear space plasma processes. They may appear as convective forms, nonlinear solitary structures, pseudo-equilibrium configurations, or other types of spatiotemporal entities. They may be locally generated or convected from elsewhere. Some of these structures may be more stable and long-lived; others may be less stable and relatively short-lived and sometimes partially formed. Generally, such structures are not purely laminar entities; they are likely composed of bundled fluctuations of all frequencies and scales. Due to the nature of the physics of complexity, it would be futile to attempt to evaluate and/or study the details and stabilities of each of these countless interacting structures or partially formed structures, though some basic understanding of the different types of structures would be helpful in the comprehension of the full stochastic complexity that results from their underlying nonlinear dynamics.

As an example of such coherent structures, let us consider the behavior of Alfvénic flux tubes in magnetized plasmas. We shall base our theoretical discussion by considering the plasma medium as a single, charge-neutral, conducting fluid moving in and interacting self-consistently with its electromagnetic environment. Because the plasmas in the space environment are generally highly rarefied, we can typically neglect the dissipative effects from particle collisions and assume the plasma is perfectly conducting and inviscid. The most elemental mathematical formulation for magnetized plasmas under such simplifying assumptions is the so-called ideal incompressible magnetohydrodynamics (MHD). The basic equations are the equations of continuity and motion, and the relevant Maxwell’s equations where the displacement current is ne-

glected for nonrelativistic motion.

$$\nabla \cdot \mathbf{v} = 0$$

(continuity equation for incompressible medium) (1)

$$\rho d\mathbf{v}/dt = \mathbf{j} \times \mathbf{B} - \nabla p$$

(equation of motion for an inviscid medium) (2)

$$\nabla \cdot \mathbf{B} = 0 \text{ (Gauss' law)} \quad (3)$$

$$\partial \mathbf{B} / \partial t + \nabla \times \mathbf{E} = 0$$

(Faraday's law) (4)

$$\mathbf{j} = \nabla \times \mathbf{B}$$

(Ampère's law neglecting the displacement current) (5)

where  $d/dt = \partial/\partial t + \mathbf{v} \cdot \nabla$  is the convective derivative,  $\rho$  is the density of the medium,  $\mathbf{v}$  is the velocity,  $\mathbf{j}$  is the current density,  $\mathbf{B}$  is the magnetic field,  $\mathbf{j} \times \mathbf{B}$  is the Lorentz force,  $p$  is the fluid pressure, and  $\mathbf{E}$  is the electric field. These equations are written in SI units with the vacuum permeability set equal to 1.

To complete the set of equations of such a formulation, we complement the above equations with the Ohm's law for a perfectly conducting medium:

$$\mathbf{E} + \mathbf{v} \times \mathbf{B} = 0 \quad (6)$$

These equations may be combined to form the following set of equations of motion and induction:

$$\rho d\mathbf{v}/dt = (\mathbf{B} \cdot \nabla) \mathbf{B} + \dots \quad (7)$$

$$d\mathbf{B}/dt = (\mathbf{B} \cdot \nabla) \mathbf{v} \quad (8)$$

where the eclipses represent the gradient of the total pressure ( $p + B^2/2$ ). Standard arguments lead Eqs. (7) and (8) to a linearized wave equation characterizing the fundamental propagation of small fluctuations of  $\mathbf{v}$  and  $\mathbf{B}$ . This wave equation admits the well-known Alfvén waves which are transverse waves to the magnetic field  $\mathbf{B}$  and propagate along the magnetic field direction. The phase velocity  $v_A$  of the Alfvén wave is  $B/\sqrt{\rho}$ , i. e., proportional to the strength of the magnetic field and inversely proportional to the square root of the density  $\rho$  of the plasma. For such waves to propagate, the operators on the right hand sides of (7) and (8) must not vanish, i. e.,  $\mathbf{B} \cdot \nabla \rightarrow i\mathbf{k} \cdot \mathbf{B} = k_{\parallel} \neq 0$ . Therefore, the propagation vector  $\mathbf{k}$  must contain a field-aligned component  $k_{\parallel}$ . When the parallel component  $k_{\parallel}$

of the propagation vector vanishes (i. e., at the resonance sites), the fluctuations are localized. Around these resonance sites (usually in the form of three-dimensional space curves), it may be shown that the fluctuations – which will try to propagate away as Alfvén waves – are held back by the background magnetic field and the plasmas medium, thereby forming the so-called Alfvénic coherent structures (which are actually domains of bundled stochastic fluctuations).

### Coarse-Grained Helicity

We now consider such magnetized domains near the Alfvénic resonance sites. For an ideal MHD system, any physically acceptable magnetic field must satisfy the solenoidal condition (Gauss' law):  $\nabla \cdot \mathbf{B} = 0$ . Also, any variation of the field away from the initial value must satisfy the constraints (4) and (6), i. e., Faraday's law and Ohm's law for infinite conductivity.

It may be easily demonstrated that these constraints are equivalent to an infinite set of integral constraints involving the helicity  $K$ , such that

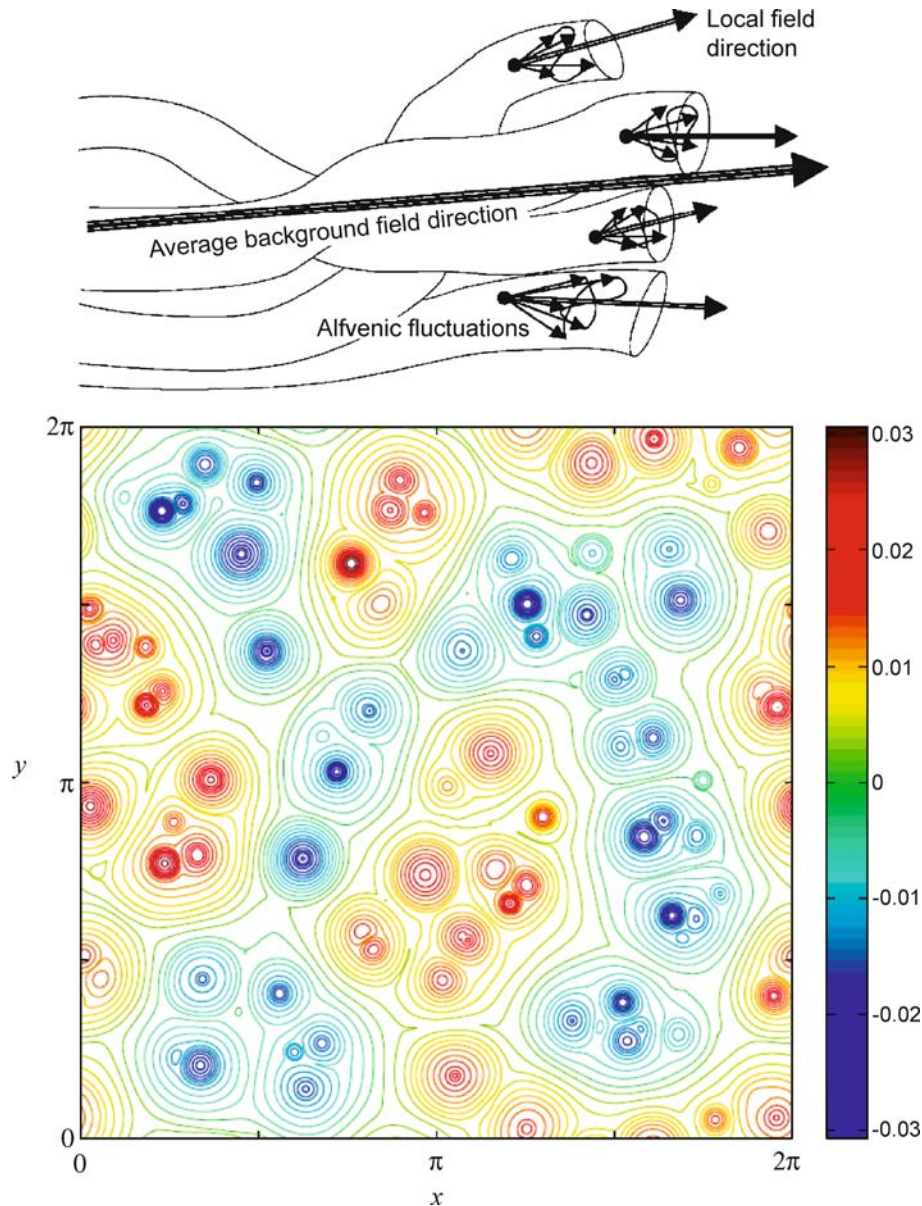
$$K = \int_V \mathbf{A} \cdot \mathbf{B} dV \quad (9)$$

is an invariant for any single connected volume  $V$  enclosed by a flux surface, where  $\mathbf{A} = \nabla \times \mathbf{B}$  is the vector potential.

### Taylor's Conjecture

We are interested in the situation where there are domains with stochastic turbulent fluctuations within which field lines merge and mix indistinguishably. Thus, it will be difficult – in fact, unpractical – to discuss the topology of individual field lines. Nonetheless, it was suggested by Taylor [114] that when the volume integral of (9) is taken over the stochastic region, the coarse-grain averaged helicity in a relaxing state would be essentially conserved. As the domain of the stochastic region near the Alfvénic resonance site relaxes to a statistically stationary minimum energy state under the constraint of this conjecture of conservation of coarse-grained helicity, it may be shown using the variational principle that such a domain (coherent structure) will be essentially force-free in the sense that  $\mathbf{j} \times \mathbf{B} = 0$  where  $\mathbf{j}$  and  $\mathbf{B}$  are the mean current density and magnetic field, respectively.

We are, of course, interested in dynamical states that are far from equilibrium. Thus, in visualizing the relaxed states from the point of view of this *Taylor's conjecture*, we shall consider timescales such that “nearly coherent” structures are formed. These structures actually move, mix



Space Plasmas, Dynamical Complexity in, Figure 3

*Top:* Schematics of tangled flux tubes in the solar wind. Each flux tube is characterized by a local magnetic field direction aligned approximately with the background field and the presence of Alfvénic fluctuations makes the magnetic field vector wander randomly about this direction. Moving across the tubes, strong intermittent interactions of predominantly nonpropagating (i.e., resonant) fluctuations are expected. Reprinted from [10] with permission from Elsevier. *Bottom:* 2D MHD simulation of cross-sectional view of interacting flux tubes for homogeneous intermittent turbulence. Colors represent current intensities and directions. Adapted from [33] with thanks to the American Institute of Physics

and sometimes merge together while immersed in an otherwise turbulently diffusing plasma medium.

To obtain some physical insight of the geometries of these magnetic coherent structures, let us consider the special situation in the solar wind and make the reasonable

assumption that the perturbed magnetic field fluctuations are much smaller than – and essentially transverse to – the mean magnetic field  $B_0$  which will be temporarily assumed to be uniform for the current discussion [89]. Thus, let us write  $\mathbf{B} = (\delta B_x, \delta B_y, B_0)$ , where  $z$  is in the direction of

the mean magnetic field, and  $(x, y)$  are orthogonal coordinates normal to  $z$ . The force-free condition for constant  $B_0$  and  $\nabla \cdot \mathbf{j} = 0$  then leads approximately to the scalar condition  $\mathbf{B} \cdot \nabla j_z = 0$ , obtained by taking the  $z$ -component of the curl of  $\mathbf{j} \times \mathbf{B} = 0$ . We have, then,

$$B_0 \partial j_z / \partial z = -(\delta B_x \partial / \partial x + \delta B_y \partial / \partial y) j_z. \quad (10)$$

For convenience, let us introduce the flux function  $\psi$  by writing  $(\partial \psi / \partial y, -\partial \psi / \partial x) = (\delta B_x, \delta B_y)$  for the perturbed transverse components of the magnetic field in the  $(x, y)$  directions such that the Gauss law of magnetism,  $\nabla \cdot \mathbf{B} = 0$  is satisfied. Then,  $j_z$  and  $\psi$  are governed by Eq. (10) and the Ampere's law,  $\nabla \times \mathbf{B} = \mathbf{j}$ .

A simple example of the flux function and axial current density satisfying the above conditions would be the class of circularly cylindrical solutions of  $\psi(r)$  and  $j_z(r)$ , [25,26]. Generally, the solutions would be more involved because of the variabilities of the local conditions of the plasma and the three-dimensional geometry. Moreover, the dynamic coherent structures with the inclusion of plasma pressure and other modifying effects (including electron-inertia terms) would be even more complicated. However, we expect these structures to be in the form of nearly field-aligned flux tubes, Fig. 3 [10,25,26,136,137,138]. Existence of Alfvénic flux tubes in the magnetopause and magnetotail have also been suggested by Tetreault [116,117] and Chang [23,24], respectively.

Generally, there exist various types of propagation modes (whistler modes, electromagnetic ion cyclotron waves, etc.) in magnetized plasmas. Thus, we envision a corresponding number of different types of plasma resonances and associated coherent structures that typically characterize the dynamics of the plasma medium under the influence of an electromagnetic background.

### Coarse-Grained Dissipation and Magnetic Reconfiguration

When coherent Alfvénic flux tubes with the same polarity migrate toward each other, strong local magnetic shears are created, Fig. 4. Wu and Chang [136,137,138] have demonstrated that the existing sporadic nonpropagating fluctuations in the strong local shear region, particularly those close to the neutral sheet (i. e., at the location where the local magnetic field vanishes), will stay in the region and continue to interact nonlinearly. On the other hand, fluctuations away from the neutral sheet region, are nonresonant and will therefore propagate away along magnetic field lines as Alfvén waves. Combined with the magnetic shear geometry, the resonant fluctuations will induce a nonlinear instability near the neutral sheet region, which

will produce more fluctuations – nonresonant and resonant. The nonresonant fluctuations will again propagate away as Alfvén waves while the resonant ones will join the other resonant fluctuations and interact nonlinearly, thereby broadening the resonance region, Fig. 5.

This combined phenomenon of “coarse-grained dissipation” depletes the energy originally contained in the coarse-grained magnetic fields near the shear region, initiating a reconfiguration of the coherent structures. In coarse-grained sense, it breaks some of the closed field lines of each of the coherent structures and then reconnects them into single closed field lines. And this process continues with the system adjusting intermittently with the surrounding environment until all free energies are exhausted, eventually leading to the formation of one single combined coherent structure with one set of coarse-grained concentric closed field lines, Fig. 6.

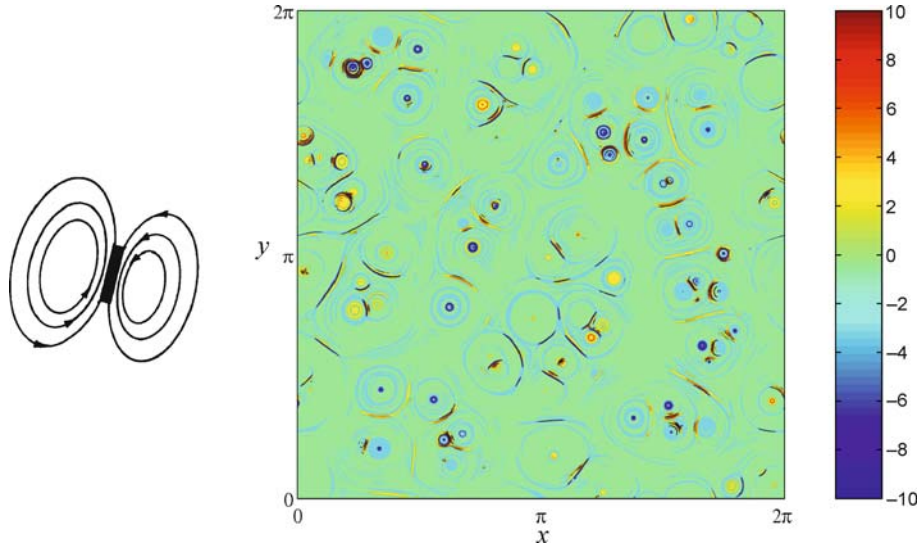
The final state of the resulting coherent structure will have less average energy due to the combined dissipation of Alfvénic propagation of nonresonant fluctuations and nonlinear interactions of the resonant fluctuations (i. e., resonance broadening). Such is the manifestation of continuous magnetic topological reconfiguration due to the dynamic “fluctuation-induced nonlinear instability”. And this merging process may repeat over and over again among the coherent structures of the same polarity, from the largest scales to the smallest scales where kinetic effects may have to be included.

On the other hand, when coherent structures of opposite polarities approach each other due to the forcing of the surrounding plasma, they might repel each other, scatter, or induce magnetically quiescent localized regions.

Under any of the conditions of the above interaction scenarios, new fluctuations will be generated. And these new fluctuations can provide new resonance sites, thereby nucleating new coherent structures of varied sizes. This kind of tangled geometry of interacting flux tubes leading to dynamical complexity has been described by Bruno et al. [10] who first deduced their existence in the solar wind, as “cooked spaghetti”, Fig. 3 (top panel), and demonstrated by Wu and Chang [136,137,138], Fig. 3 (bottom panel), Matthaeus et al. [90], Fig. 7, and others with MHD simulations. And they are the source of the observed sporadic and localized current disruptions in the Earth's magnetotail and elsewhere [75,78,82].

Topological reconfigurations of such nature occur quite frequently for the dynamical interactions of coherent structures in space plasmas and are not limited just to flux tubes; for example, Sundkvist et al. [109] and Alexandrova et al. [1] have observed intermittent interactions of drift Alfvén vortices in the cusp and magnetosheath, respec-





**Space Plasmas, Dynamical Complexity in, Figure 4**

Cross-sectional views of interacting Alfvénic flux tubes. *Left:* Schematic of merging. Arrows indicate directions of magnetic field and blackened area indicates location of strong shear. *Right:* 2D MHD simulation of the current intensities for homogeneous intermittent turbulence. Colors represent the intensities and directions. Adapted from [137] with permission from Elsevier

tively. Similarly, such enhanced intermittency at the intersection regions of whistler coherent structures has also been surmised by Consolini and Lui [40] and Consolini et al. [42] in the plasma sheet. All such intermittent interactions are quite akin to the avalanche phenomenon prevalent in sandpile models [3,35,36,62,133]. They are the origin of the various observed magnetic reconnection signatures in space plasmas.

This stochastic behavior of the interactions of the plasma coherent structures is a phenomenon of “dynamical complexity” as defined in Sect. “Definition of the Subject”, which, for emphasis, is reproduced below:

*“Dynamical complexity” is a phenomenon exhibited by a nonlinearly interacting dynamical system within which multitudes of different sizes of large scale “coherent structures” are formed, resulting in a complicated global nonlinear stochastic behavior for the dynamical system – vastly different from that could be surmised from the original dynamical equations.*

### Non-Gaussian Probability Distribution Functions

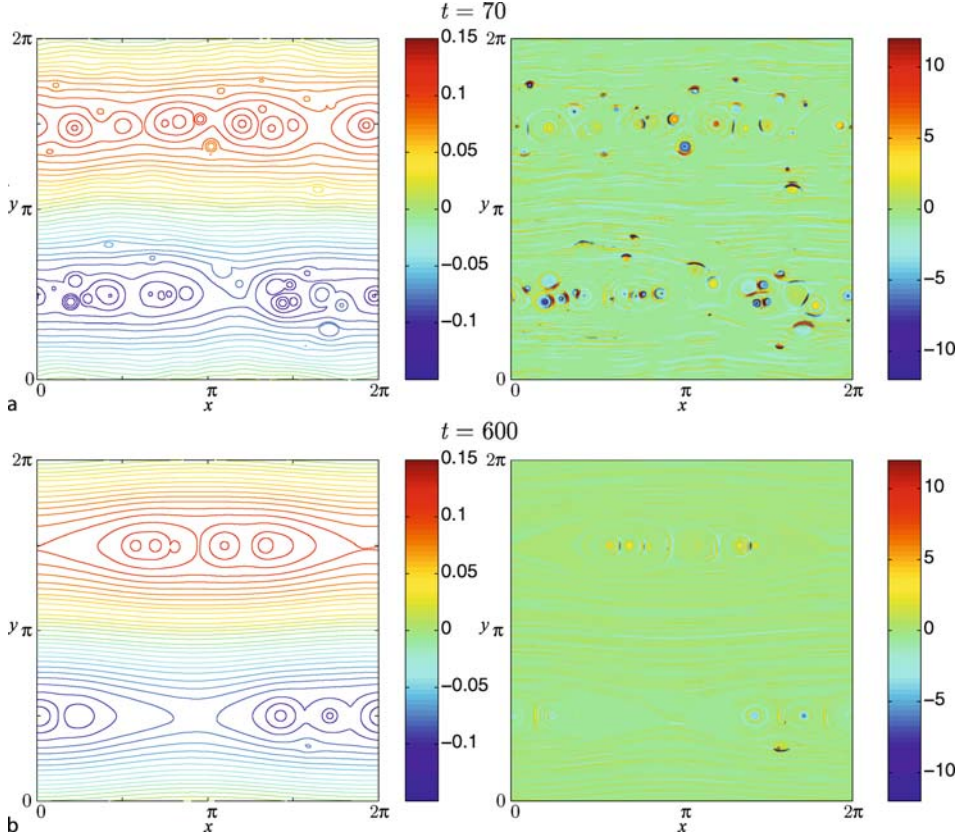
The fluctuations that are induced by the interactions and mergings of coherent structures are sporadic and localized. Since the coherent structures are numerous and out-sized, we expect the fluctuations within the interaction regions of these structures (resonance overlap regions) to be large and occur relatively more frequently than those that

would have been expected from a medium of the original minute plasma particles (electrons, ions and neutrals). A technique useful in gauging the degree of such effects is by studying the shapes of the probability distribution functions (PDFs) of intermittent fluctuations at varying scales.

To demonstrate this, let us refer to one of the 2D MHD simulations described previously [32,33,136,137,138]. (In the following discussions, the measured fluctuations may be those of any physical property of the MHD medium. We shall choose the physical property as the strength of the magnetic field  $B(x_i)$  to render the discussions more specific). Consider, for example, the spatial series of the strength of the magnetic field  $B(x_i)$  for a given time  $t$  and at fixed  $y$ , with  $x_i = i\delta$  where  $i = 0, 1, 2, \dots, N$ , and  $\delta$  is the grid size of the simulation. We can then form the coarse-grained differences

$$\delta B_i^2(\Delta) = B^2(x_i + \Delta) - B^2(x_i) \quad (11)$$

within the interval with  $\Delta = k\delta$  ( $k$ : an integer) and generate the probability distribution function  $P(\delta B^2, \Delta)$ . Figure 8 displays the calculated results of  $P(\delta B^2, \Delta)$  for a 2D MHD simulation for the homogeneous case for several coarse-grained scales,  $\Delta$ . From this figure, we note that the distribution for each scale falls nearly onto a smooth curve with the exception at the tails where scatterings from the mean are more visible. And the shapes of the PDF curves deviate more and more from that of a Gaussian at smaller and smaller scales.



Space Plasmas, Dynamical Complexity in, Figure 5

2D MHD simulation of coarse-grained dissipation in sheared magnetic field. *Left:* Contours of magnetic flux. *Right:* Corresponding current density distributions. Initially, new fluctuations are excited and the resonant fluctuations begin to interact nonlinearly near the neutral sheet region (near  $y = \pi/2, 3\pi/2$ ). Eventually, the field lines are reconnected in the coarse-grained sense. Adapted from [137] with permission from Elsevier

Such PDFs sometimes satisfy the one-parameter scaling form:

$$P(\delta B^2, \Delta)\Delta^s = P_s(\delta B^2/\Delta^s) \quad (12)$$

where  $s$  is the parameter (or scaling exponent) such that all PDFs essentially collapse onto one master scaling function  $P_s$ . Such scaling behavior seems to be approximately satisfied for the above simulated result with  $s \approx 0.335$  (Fig. 9), although closer examination reveals some discrepancies – particularly in the tail regions [33]. The more subtle nature of the scaling properties of such PDFs will be considered further in Sect. “Multifractals”.

We shall now proceed to make contact with in-situ spacecraft observations. Until recently, most of the turbulence data in space came from measurements obtained by one single spacecraft. The data are then analyzed based on the Taylor hypothesis [113] which assumes that the transit time of eddies (bundled fluctuations) is much less

than the characteristic evolution time. For fully developed MHD turbulence, the characteristic eddy turnover (evolution) time may be estimated:

$$t(\text{evolution}) \sim (\lambda/2\pi)/[(\delta B/B_0)v_A] \quad (13)$$

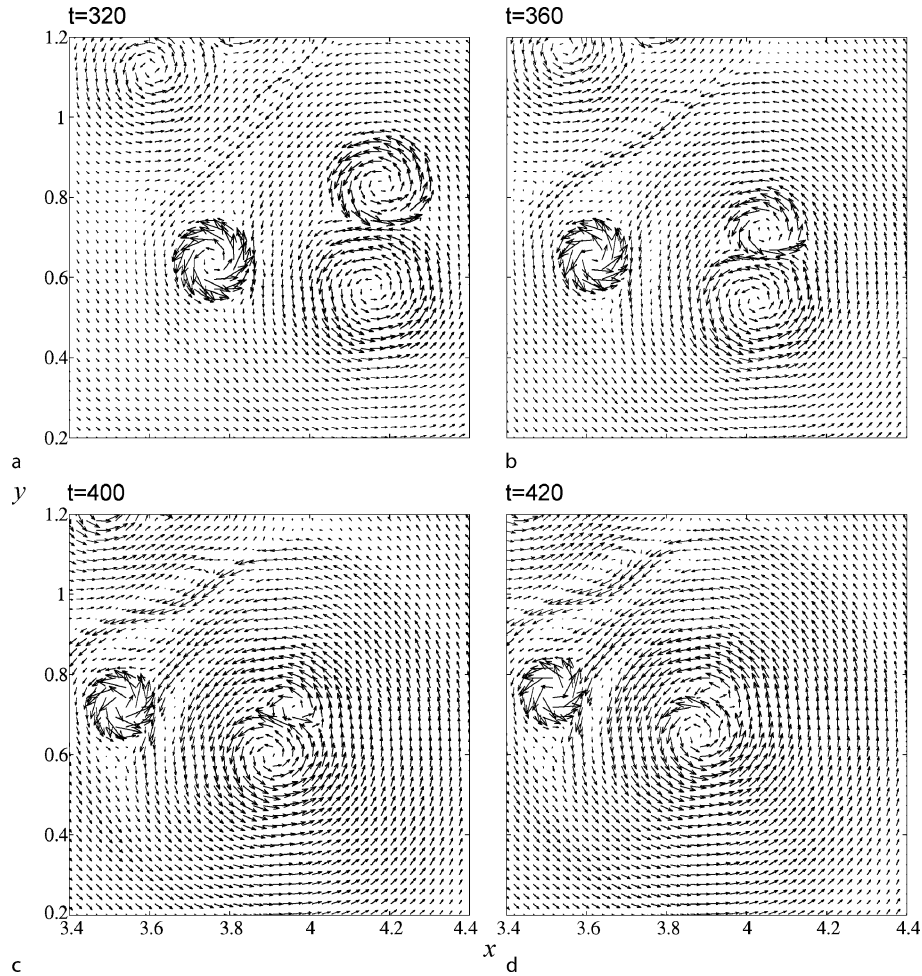
where  $\lambda$  is the typical size of the eddy,  $v_A$  is the Alfvén speed based on the mean magnetic field  $B_0$  and  $\delta B$  is the average fluctuating magnetic field of the eddy in question. The transit time of the eddy is:

$$t(\text{transit}) \sim \lambda/v_s \quad (14)$$

where  $v_s$  is the relative streaming speed between the plasma medium and the spacecraft.

Thus, as it has been demonstrated by Matthaeus and Goldstein [88] that the requirement of  $t(\text{evolution}) \gg t(\text{transit})$  becomes:

$$v_s/v_A \gg 2\pi(\delta B/B_0) \quad (15)$$



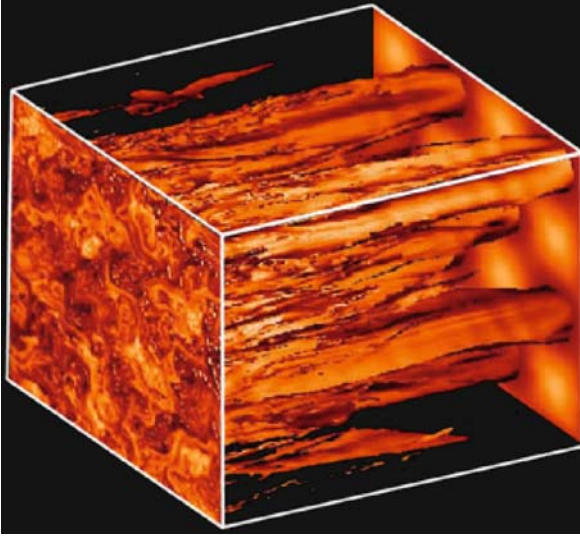
Space Plasmas, Dynamical Complexity in, Figure 6

2D MHD simulation of merging of flux tubes. Arrows indicate magnitudes and directions of magnetic fields. Reprinted from [137] with permission from Elsevier

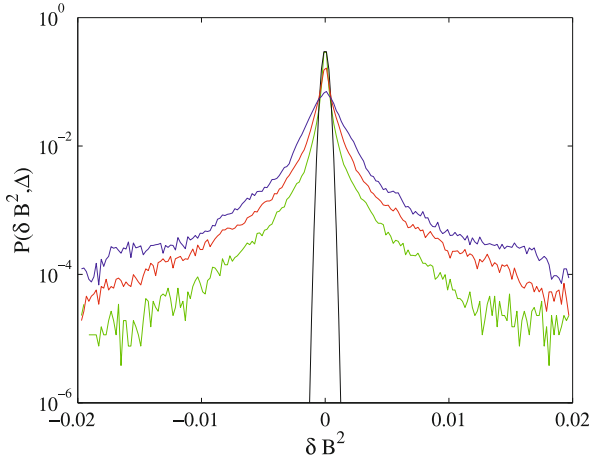
As pointed out by Tu and Marsch [120], at 1 AU for solar wind observations the wind speed (which is essentially the relative speed between the plasma medium and the spacecraft since it is much larger than the spacecraft speed) is of the order of 10 times that of the ambient Alfvén speed and  $\delta B/B_0 \simeq 0.5$ . Therefore, condition (15) is satisfied for solar wind turbulence in that region. Because the spacecraft speed is much smaller than the solar wind speed, the data gathered by instruments aboard the spacecraft essentially sample the spatial fluctuations in the direction of the solar wind, particularly at times when the solar wind turbulence is homogeneous at large scales and fully developed. Thus, for example, the time series of such measurements of the strength of the magnetic field  $B(t_i)$ , where  $t_i = i\delta$  with  $i = 0, 1, 2, \dots, N$  and  $\delta$  is the sampling interval, may be interpreted based on the hypothesis as the spatial series

of the ambient turbulence fluctuations with  $x = v_s t$  ( $x$ : direction of the solar wind velocity). We can then construct the probability distribution function  $P(\delta B, \tau)$  or  $P(\delta B^2, \tau)$  with the coarse-grained scale  $\tau = k\delta$  ( $k$ : integer) in analogy to the calculations related to the simulation results discussed above.

Non-Gaussian probability distribution functions from such statistics have been commonly observed in the solar wind for fluctuations related to the magnetic, velocity, density and other field variables, e.g. [15,47,48,56,71,72,85,107,108]. Figure 9 depicts the typical PDFs of magnetic field fluctuations obtained from such an analysis for both the slow ( $\sim 400 \pm 200$  km/s) and fast ( $\sim 700 \pm 100$  km/s) solar wind [108]. The similarities between the simulated results and results from such observations as exhibited in Fig. 9 are quite striking.

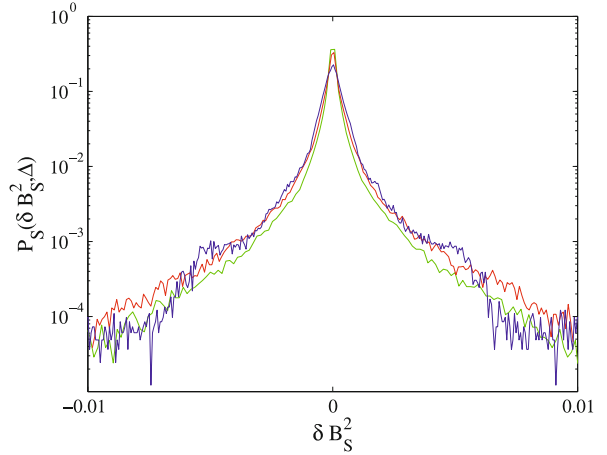


**Space Plasmas, Dynamical Complexity in, Figure 7**  
MHD simulation of a two-component model, with 80% of its energy in two-dimensional modes, exhibiting tangled flux tubes with considerable transverse complexity. Reprinted from [90] with permission from the American Astronomical Society



**Space Plasmas, Dynamical Complexity in, Figure 8**  
Probability distribution function  $P(\delta B^2, \Delta)$  from a 2D MHD simulation for homogeneous turbulence for  $\Delta = 2$  (green), 8 (red), 32 (blue) units of grid spacing. The black curve is the PDF for Gaussian fluctuations. Reprinted from [34] with permission from Springer Science and Business Media

One popular empirical technique in expressing the shape of the PDFs of the observed intermittent turbulent fluctuations is the Castaing distribution. In 1990, Castaing et al. [21] suggested that the intermittent fluctuations might perhaps be viewed as an ensemble of subsets of fluctuations  $\xi$ , each subset having a normal distribution:



**Space Plasmas, Dynamical Complexity in, Figure 9**  
Scaled probability distribution function  $P_s(\delta B_s^2, \Delta)$  with  $\delta B_s^2 = \delta B^2 / \Delta^s$ ,  $s = 0.335$ , and  $\Delta = 2$  (green), 8 (red), 32 (blue) units of grid spacing. Adapted from [33] with thanks to the American Institute of Physics

$$P_\sigma(\xi) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\xi^2/2\sigma^2) \quad (16)$$

where  $\sigma$  is the variance. The intermittency is then assumed to be due to the fluctuations of the variances that satisfy a log-normal distribution:

$$Q_\lambda(\sigma)d\sigma = \frac{1}{\lambda\sqrt{2\pi}} \exp\left(-\frac{\ell n^2(\sigma/\sigma_0)}{2\lambda^2}\right) d\ell n\sigma \quad (17)$$

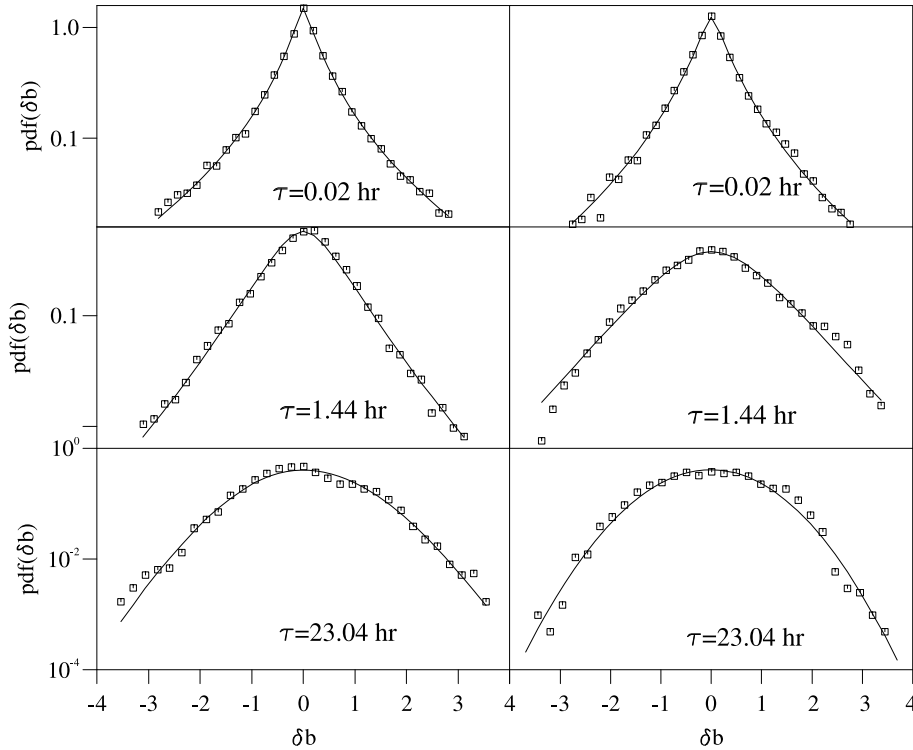
where  $\sigma_0$  is the most probable variance of  $\xi$ , and  $\lambda$  is the variance of  $\ell n\sigma$ . Combining (16) and (17) gives the Castaing distribution

$$\Pi_\lambda(\xi) = \frac{1}{2\pi\lambda} \int_0^{+\infty} \exp\left(-\frac{\xi^2}{2\sigma^2}\right) \exp\left(-\frac{\ell n^2(\sigma/\sigma_0)}{2\lambda^2}\right) \frac{d\sigma}{\sigma^2}. \quad (18)$$

Such fits of data to (18) are shown as solid curves in Fig. 10. How well does such an ansatz characterize the data depends very much on the type and scale of the fluctuations [47]. Other empirical fits such as the kappa-distribution and gamma-distribution have also been suggested, but we do not have space here to enumerate and discuss these and other phenomenological models.

Similar observational results have been detected for the magnetic field and velocity fluctuations in the plasma sheet [2,134], in the cusp [44,139], for the electric field





**Space Plasmas, Dynamical Complexity in, Figure 10**

PDF for  $\delta B$  calculated from the observational data in the (left) fast solar wind ( $\sim 700$  km/s) and (right) slow solar wind ( $\sim 400$  km/s) streams. Solid lines represent fits obtained with the model suggested by Castaing et al. [21]. Reprinted from [108] with permission from Elsevier

fluctuations in the auroral zone [111,112], as well as for fluctuations of other field variables, though caution must be exercised in interpreting some of these results as the Taylor hypothesis may or may not be satisfied in the observational regions.

### Wavelet Transforms and Intermittency

Traditional procedure for analyzing fluctuations is to perform Fourier transforms of the observational, numerical or experimental data. The basis functions of the Fourier transform are sines and cosines. Each such basis function extends in space (or time) all the way from minus infinity to plus infinity. Thus, Fourier transform essentially wipes out the localized information of the intermittent fluctuations resulting from the confined interactions of the coherent structures. Therefore, for intermittent turbulence, it will be more appropriate to consider transforms whose basis functions are localized and scale dependent instead of sines and cosines. Such a transform has the generic name of “wavelet transform” [43,46]. Below, we restrict our discussions to one-dimensional wavelet transforms.

To be more specific, a wavelet transform replaces (transforms) the fluctuating spatial (or temporal) series into the coefficients of a set of basis functions that are localized in space (or time) with different scales in space (or time). The basis functions are usually chosen such that they are square-integrable and generated by a single mother wavelet, a rescalable and translatable function  $\psi[(x - b)/a]$  where  $a$  is the scale and  $b$  defines the translation along  $x$ . Two of the more popular mother wavelets are the complex-valued Morlet wavelet [94] defined by

$$\psi_0(y) = e^{ik \cdot y} e^{-|y|^2/2} \quad (19)$$

and the Haar wavelet [52] given by

$$\psi_0(y) = \begin{cases} 1 & \text{if } y \in [0, 1/2); \\ -1 & \text{if } y \in [1/2, 1), \text{ and} \\ 0 & \text{if } y \notin [0, 1). \end{cases} \quad (20)$$

To generate the full set of functions, we simply replace the argument in (19) or (20) by  $(x - b)/a$  and allow the parameters  $(a, b)$  to vary within the domain of interest.



The Morlet wavelet is appealing to the physicists because it fits in with the usual understanding of wave packets. And due to its continuous nature, the transform is amenable to analytical manipulations. One disadvantage of the Morlet wavelets is that they are not entirely orthogonal to each other. The Haar wavelet, on the other hand, is truly local and the basis functions are real and orthogonal to each other. The simple and discrete nature of the Haar wavelets renders it ideal for numerical applications. One disadvantage of the Haar wavelets is that they are not symmetrical with respect to the midpoint, thus producing slanted spectral plots. This asymmetry, however, becomes less and less noticeable for smaller and smaller scales.

To obtain the coefficients of a wavelet transform for a particular fluctuating spatial (or temporal) series  $S(x)$ , we evaluate the convolution integral as follows:

$$C(a, b) = (K/a) \int S(x') \psi[(x' - b)/a] dx' \quad (21)$$

where  $K$  is a normalization constant.

From  $C(a, b)$ , we can calculate, for example, the normalized power  $NP(a, b)$ :

$$NP(a, b) = |C(a, b)|^2 / \langle |C(a, b)|^2 \rangle_x \quad (22)$$

where  $\langle \dots \rangle_x$  denotes the spatial (or temporal) average; and, here, the average is essentially the Fourier transformed power for the entirety of the fluctuations. In usual applications, only a truncated finite set of coefficients are used to perform practical calculations. The normalized power is a power spectrum of both scale and location; thus, it is sometimes called the Local Intermittency Measure, LIM. By definition, LIM is equal to 1 for the Fourier transform. Figure 11(top) is a color representation of the Local Intermittency Measure or LIM of the magnetic field intensity using the Haar transform for the 2D homogeneous MHD simulation discussed previously at a given time  $t$  and for fixed  $y$ . The scale-dependent, sporadic and localized powers of the fluctuations are vividly displayed. We note that the intermittency increases as the scale is reduced. And, the intensity can be rather strong at small scales – even when the fluctuations are strongly intermittent and localized.

Normalized power of wavelet transforms have been used to evaluate the Local Intermittency Measure LIM in the solar wind (e.g. [9,10]), for the AE index [38,39], in the magnetotail (e.g. [42,77,131]), in the cusp (e.g. [44,139]), and in the auroral zone [112]. Figure 11 (bottom) is a color representation of the Local Intermittency Measure

LIM using the Haar transform for the electrical field fluctuations in the transverse direction to the magnetic field in the auroral zone [112]. The characteristics of strong intermittency at small scales are clearly visible.

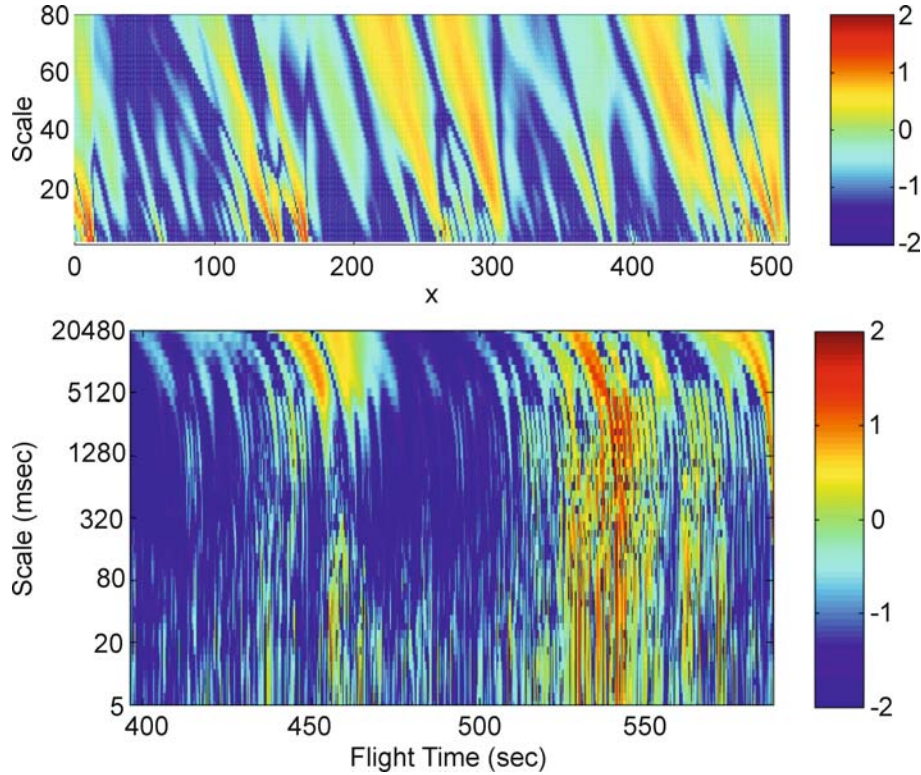
As it has been pointed out in Sect. “Non-Gaussian Probability Distribution Functions”, intermittent fluctuations have non-Gaussian probability distributions especially at small scales. We may look for a quantity which provides a measure of the non-Gaussian nature of the fluctuations as a function of scale. A normalized scale dependent measure related to the fourth order moment of the coarse-grained probability distribution is called the Flatness. It indicates whether the data are more peaked or flatter relative to the Gaussian distribution. It is also called the Kurtosis. It has been suggested by Meneveau [91] that the average over space (or time) of the square of the normalized power is essentially the scale dependent Flatness of the probability distribution.

$$\text{Flatness} = \langle [NP(a, b)]^2 \rangle_x \quad (23)$$

For Gaussian probability distributions, the value of the Flatness may be shown to be equal to 3. For an intermittent event, the Flatness increases without bound at small scales. The rate at which the Flatness increases with the decrease of scale is a measure of intermittency. Figure 12(top) plots the values of the Flatness calculated as a function of scale corresponding to the fluctuations discussed above for the magnetic field fluctuations for the 2D MHD simulation for given  $y$  and fixed  $t$ . We note that the value of the Flatness becomes larger than 3 and continues to increase as the scale becomes smaller and smaller.

Flatness has been evaluated for the solar wind turbulence by Bruno et al. [10,11], Fig. 12(bottom). It is one of the tools used by Bruno et al. [10] to arrive at the conclusion that solar wind turbulence is the manifestation of interacting tangled flux tubes or “cooked spaghetti” as depicted in Fig. 3(top). Wavelet transforms have been used by various authors to evaluate Flatnesses of intermittent turbulence in other regions of space plasmas, e.g. [44,77,112,131,139]. Other more generalized techniques of gauging the intermittency in space plasma turbulence have also been employed by Vörös [129] and Vörös et al. [130,132], which we shall not discuss here.

The normalized power and Flatness are only two tools that provide gauges of intermittency of a fluctuating series. One might ask: Are there other measures such that in their totality fully characterize the intermittency nature of the fluctuating series? This will be addressed briefly in the next section.



Space Plasmas, Dynamical Complexity in, Figure 11

**Top:** Normalized power or LIM for the intermittent magnetic field fluctuations  $B(x)$  using the Haar transform for a 2D homogeneous MHD simulation at given time  $t$  and for fixed  $y$ , in units of grid spacing. *Color* indicates intensity. **Bottom:** Normalized power or LIM for the intermittent transverse electric field fluctuations in the auroral zone. *Color* indicates intensity. Adapted from [112] with thanks to the American Geophysical Union

### Multifractals

Another popular *modus operandi* designed to study the phenomenon of intermittency is based on the concept of multifractals. As we have seen above, turbulence in space plasmas generally encompass fluctuations of all varieties and sizes, which interact and propagate throughout the plasma medium. For illustrative purposes, let us visualize some particular fluctuations that have conventional geometrical properties in a three-dimensional Euclidean space. Because of their sporadic and localized nature, it is easy to imagine that they generally cannot fill the full three-dimensional space that they occupy at a given time. Or said in another way, the space these fluctuations occupy is only a fraction of the full three-dimensional space. Such geometrical property was popularized by Mandelbrot [83] when he coined the word “fractals” or fractal geometry.

Actual fluctuations in plasma turbulence generally do not have the conventional geometrical properties. We must then devise some abstract “measure” to characterize the properties of the fluctuations and evaluate its frac-

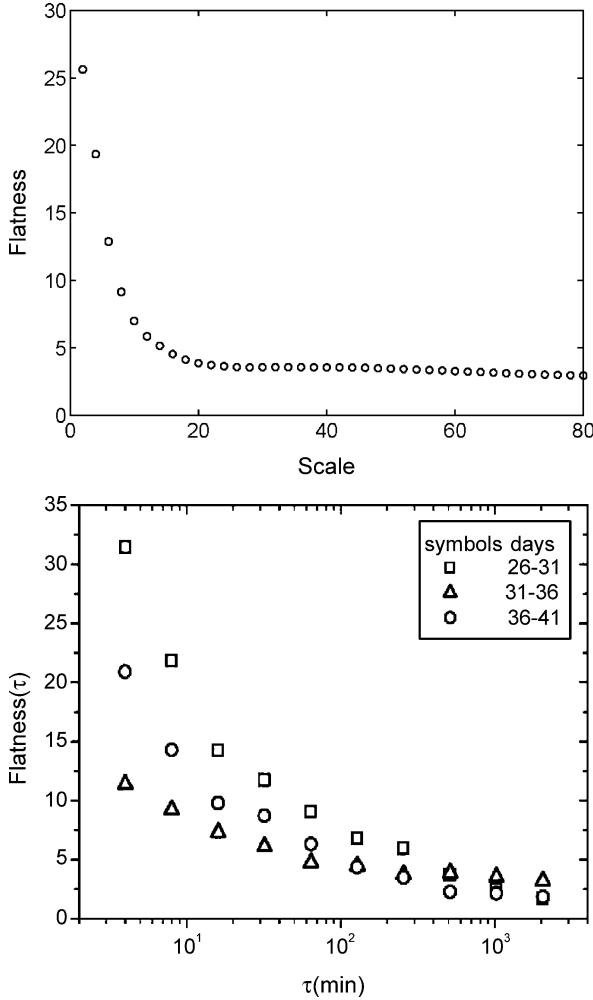
tal characteristics which may be interpreted with geometrical analogs. Consider, for example, the spatial series of the simulated fluctuations of the strength of the magnetic field,  $B(x_i)$ , along some constant value of  $y$  at time  $t$  for the two-dimensional homogeneous MHD turbulence discussed in the previous sections, where  $x_i = i\delta$  with  $i = 0, 1, 2, \dots, N$  and  $\delta$  is the length between grid points. From this series, we can construct a spatial series by considering, for example, the absolute value of the fluctuations due to the coarse-grained difference of the strength of the magnetic field between two spatial values  $x_i + \Delta$  and  $x_i$  with  $\Delta = k\delta$ , some multiple of  $\delta$ :

$$\delta B_i = |B(x_i + \Delta) - B(x_i)| \quad (24)$$

within the spatial interval  $X = N\delta$ . The simulation is statistically homogeneous over the interval  $X$ . Thus, we may calculate the ensemble average of  $\delta B_i$  over the interval  $X$ ,

$$S_1(\delta B, \Delta) = \langle |B(x_i + \Delta) - B(x_i)| \rangle \quad (25)$$

and use it as a “measure” for the coarse-grained fluctuations of the simulated spatial series. The choice of taking



Space Plasmas, Dynamical Complexity in, Figure 12

**Top:** Flatness as a function of scale in units of grid spacing for the intermittent magnetic field fluctuations  $B(x)$  using the Haar transform for a 2D homogeneous MHD simulation at given time  $t$  and for fixed  $y$ . **Bottom:** Flatness versus time scale for three different time intervals for the solar wind during solar minimum (i. e., when the solar activity is very weak) as measured by the WIND spacecraft. Reprinted from [11] with thanks to the American Institute of Physics

the ensemble average of the absolute values of the coarse-grained differences instead of the values of the raw differences is for the purpose of better statistical convergence [8, 127]. We can then plot  $\log S_1(\delta B, \Delta)$  against  $\log \Delta$  for different choices of  $\Delta$  (or  $k$ ). If the result approximates a straight line for some range of  $\Delta$  for small  $\Delta$ , we can then assign a fractal number  $\zeta_1(\delta B)$  to the fluctuations of the strength of the magnetic field as:

$$\zeta_1(\delta B) = d(\log S_1(\delta B, \Delta))/d(\log \Delta) \quad (26)$$

for this range of small  $\Delta$ . In other words, within this range of small  $\Delta$  we may represent  $S_1(\delta B, \Delta)$  as power of  $\Delta$  with exponent  $\zeta_1(\delta B)$ . This exponent may be considered as an analog to the classical concept of dimension. It is generally an irrational number and usually cannot be surmised from simple dimensional analysis arguments; thus it is sort of a “fractal dimension” for the particular choice of “measure” described above.

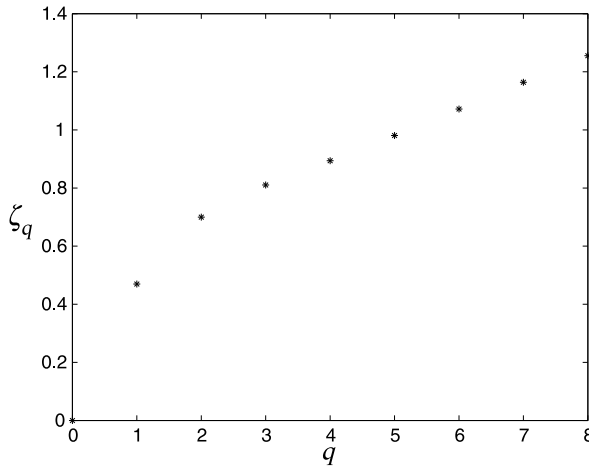
### Structure Functions

It is obvious that if another choice of measure is made, the corresponding fractal number for the same spatial series will generally be different. For example, one might wish to look at higher order moments of the PDFs  $P(\delta B, \Delta)$  of  $\delta B$ , the so-called structure functions:

$$\begin{aligned} S_q(\delta B, \Delta) &= \int (\delta B)^q P(\delta B, \Delta) d(\delta B) \\ &= \langle |B(x_i + \Delta) - B(x_i)|^q \rangle \end{aligned} \quad (27)$$

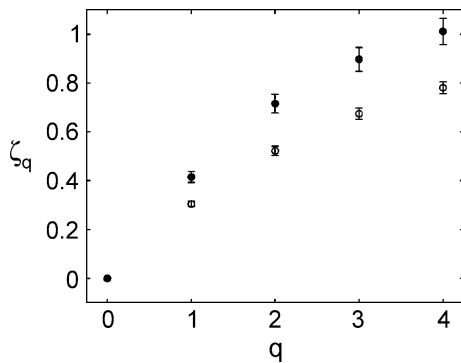
The motivation here is that different moments emphasize different peaks in the fluctuating series. Generally, corresponding to each  $\zeta_q$  there will be a fractal exponent  $\zeta_q$  for small  $\Delta$ . If  $\zeta_q = \zeta_1 q$ , then the fractal property of the fluctuating series is fully characterized by the value of  $\zeta_1$ . Such fluctuations are then said to be “monofractal” or “self-similar”, i. e., the fractal characteristics for all the moment orders are similar to each other. For general intermittent turbulence, on the other hand,  $\zeta_q$  would be a non-linear function of  $q$ , or “multifractal”.

In practical calculations, it has been suggested by Benzi et al. [6] for hydrodynamic turbulence studies that the correlations among the structure functions seemed to be much more apparent than those exhibited by the correlations of individual structure functions directly with coarse-grained scaling. This is particularly true for structure functions based on the absolute values of the differences of fluctuations as it is defined here [51]. This also seemed to be true for plasma turbulence studies [19, 20, 95, 100]. Thus, it has become very popular to first determine the relative values of the structure function exponents with respect to the structure function exponent of a particular order, and then determine the exponent for that particular order more accurately by some phenomenological, dimensional, or empirical reasoning. This procedure is called “extended similarity scaling” (ESS). Figure 13 is the result of such a statistical calculation for the fluctuations of the magnetic field strength  $B(x_i)$  of the 2D homogeneous MHD simulation for given  $y$  and  $t$ . We note that the structure function exponents exhibit nonlinear, i. e., multifractal, behavior as expected for intermittent turbulence.



Space Plasmas, Dynamical Complexity in, Figure 13

Structure function exponents for the magnetic field intensity at given time  $t$  and for fixed  $y$  using ESS for a 2D MHD simulation for homogeneous intermittent turbulence. The values of the exponents at high orders are less accurate than those for the low orders due to the size limitation of the simulation box. Reprinted from [27] with thanks to the American Institute of Physics



Space Plasmas, Dynamical Complexity in, Figure 14

Structure function exponents using ESS for the time series of the strength of the magnetic field collected for extended periods during solar minimum (low solar activity) by the WIND spacecraft (circles) and solar maximum (high solar activity) by the ACE spacecraft (filled circles). Adapted from [11] with thanks to the American Institute of Physics

Calculations for structure function exponents have been carried out for magnetic and velocity fluctuations in the solar wind, e. g. [11,12,13,14,17,58,59,60,86,97,121] and elsewhere in the space environment, e. g. [57], usually by assuming the applicability of the Taylor hypothesis. We caution here again that except for most of the solar wind studies, the assumption of the Taylor hypothesis must be exercised with a critical eye. In Fig. 14, the results for such calculations for the time series of the strength of the mag-

netic field collected for extended periods during solar minimum (low activity) and solar maximum (high activity) by two separate spacecraft (Wind and ACE) are presented. The multifractal nature of the intermittent turbulence is apparent in both calculated results [11].

Structure function calculations may be performed conveniently for a fluctuating series for positive values of  $q$ ; but will generally diverge for  $q < 0$ . An alternative procedure to evaluate the fractal characteristics of different moment orders is the so-called “singularity analysis” described below.

### Partition Functions, Generalized Dimensions and Singularity Spectra

As we have seen in the previous discussions, turbulence in space plasmas is generally intermittent and therefore probably composed of an admixture of fluctuations of different fractal characteristics. We may attempt to extract this “multifractal” nature of the observed fluctuating time series by searching for the “dominant singular behavior” for different moment orders at small scales following the technique suggested by Parisi and Frisch [98] and Halsey et al. [54] and first applied to solar wind turbulence by Burlaga [12,13,14,15], and later for the auroral electrojet (AE) index by Consolini [41], for the high latitude geomagnetic fluctuations by Vörös [129,130], for the plasma sheet by Lui [76] and Weygand et al. [134], to the magnetospheric cusp by Yordanova et al. [139], and to the auroral electric field fluctuations by Chang et al. [34].

The basic idea here is to define an appropriate measure that is scale dependent. (Instead of referring to a spatial series such as that considered above for the numerical simulation, for variance, we consider below a generic time series based on temporal observations). To proceed, we first define an “incremental measure”:

$$\delta\mu_j = |B(t_j + \delta) - B(t_j)| \quad (28)$$

where  $\delta$  is the sampling interval. We now subdivide the total time interval  $T$  into  $M = T/\tau$  segments with  $\tau = k\delta$  and calculate the normalized scale-dependent “segmental measure”

$$\mu_i(\tau) = \sum_{j=(i-1)k+1}^{ik} \delta\mu_j / \mu \quad \text{with} \quad \mu = \sum_{j=1}^N \delta\mu_j \quad (29)$$

For convenience, we will generally choose  $M$  to be an integer.

The normalization of  $\mu_i(\tau)$  is an attempt to represent the segmental measure as a measure of probability. This

is the scale dependent measure that we seek to define. We shall assume that each such measure varies with the scale  $\tau$  in a singular manner as a power law. We now form the  $q$ th moment order of the coarse-grained probabilities  $\mu_i(\tau)$ , traditionally called the “partition function” [54]:

$$\Gamma(q, \tau) = \sum_{i=1}^M \mu_i^q(\tau) \quad (30)$$

where  $q$  can be any real number. We shall now search for the dominant singular behavior of  $\Gamma(q, \tau)$  as characterized by a power law in  $\tau$  with exponent  $\gamma(q)$  for small  $\tau$  similar to that for the structure function analysis. In general, for each moment order,  $\gamma(q)$  is a different number characterizing the particular fractal behavior of the subset of fluctuations, which dominates the (singular) scaling behavior of that particular moment order.

As mentioned above, Consolini et al. [41] have carried out such a singularity measure multifractal analysis for a typical time series of the auroral electrojet (AE) index, Fig. 15 (top left panel). In their calculations, the incremental measure was chosen as:

$$\delta\mu_j = |\phi_{\text{AE}}(t_j + \delta) - \phi_{\text{AE}}(t_j)|^2 \quad (31)$$

where  $\phi_{\text{AE}}(t_j)$  is the measured AE index at time  $t_j$  and  $\delta$  is the sampling interval. We note from the figure that  $\gamma(q)$  is indeed a nonlinear function of  $q$ , and thus, a “multifractal”. There are two distinct asymptotic regimes of  $q$  ( $q < -2$ , and  $q > +2$ ) for which  $\gamma(q)$  varies with  $q$  nearly linearly and a nonlinear crossover region between the two asymptotes.

Henschel and Procaccia [55] suggested the concept of the so-called “generalized dimensions” for fractals. Their definition of the generalized dimension  $D_q$  is related to the partition function  $\Gamma(q)$  as follows:

$$D_q = (q-1)^{-1} \lim_{\tau \rightarrow 0} \log \Gamma(q) / \log \tau = \gamma(q)/(q-1). \quad (32)$$

It is called the generalized dimension because in the limit of  $q \rightarrow 0$ ,  $D_0$  is essentially the fractal (or similarity) dimension defined by Mandelbrot [83]. Also, it may be shown that, in the limit of  $q \rightarrow 1$ ,  $D_1$  corresponds to the so-called information dimension [5]. In addition,  $D_q$ 's for  $q \geq 2$  are the corresponding  $q$ th correlation dimensions [50]. The calculated generalized dimension  $D_q$  as a function of moment order  $q$  for the AE index is shown in Fig. 15 (top right panel). As shown by Consolini et al. [41], such a nonlinear variation may be modeled rather accurately by the so-called  $p$ -model which is a “two-scale cascade model” as fol-

lows [16,92]:

$$D_q = \log_2[p^q + (1-p)^q]^{1/(1-q)} \quad \text{with } p = 0.746 \pm 0.002. \quad (33)$$

where  $p$  is the single fitting parameter associated with the fragmentation probability of the cascade process. This one-parameter representation of the generalized dimension for the multifractal phenomenon of the AE index reinforces the idea stated earlier that the multifractal phenomenon can be thought of as the crossover phenomenon between two invariant regions.

Another way to gauge the multifractal characteristics in terms of the partition function formalism is the method of the singularity spectrum  $f(\alpha)$  [54,98]. If we associate with each segmental measure  $\mu_i$  a singularity index  $\alpha_i$  such that  $\mu_i = \tau^{\alpha_i}$  where  $\alpha_i$  is within some range between  $\alpha$  and  $\alpha + \Delta\alpha$  and pass on to the continuum limit by writing the number of differential measures whose singularity index is between  $\alpha$  and  $\alpha + d\alpha$  as  $\rho(\alpha)\tau^{-f(\alpha)}d\alpha$ , then the partition function as defined in (30) in the continuum limit is:

$$\Gamma(q, \tau) = \int d\alpha' \rho(\alpha') \tau^{-f(\alpha') + q\alpha'}. \quad (34)$$

Since we are interested in the phenomenon for very small  $\tau$ , the integral is dominated by the value of  $\alpha'$  which makes the expression  $-f(\alpha') + q\alpha'$  the smallest. Assuming  $\rho(\alpha') \neq 0$  and  $f(\alpha')$  smooth, the minimum is located at  $\alpha' = \alpha(q)$  with

$$f'(\alpha(q)) = q \quad \text{and} \quad f''(\alpha(q)) < 0 \quad (35)$$

Therefore the partition function for very small  $\tau$  is approximately

$$\Gamma(q, \tau) \sim \exp \{ [q\alpha(q) - f(\alpha(q))] \ln \tau \} \quad (36)$$

This leads immediately to the identification of

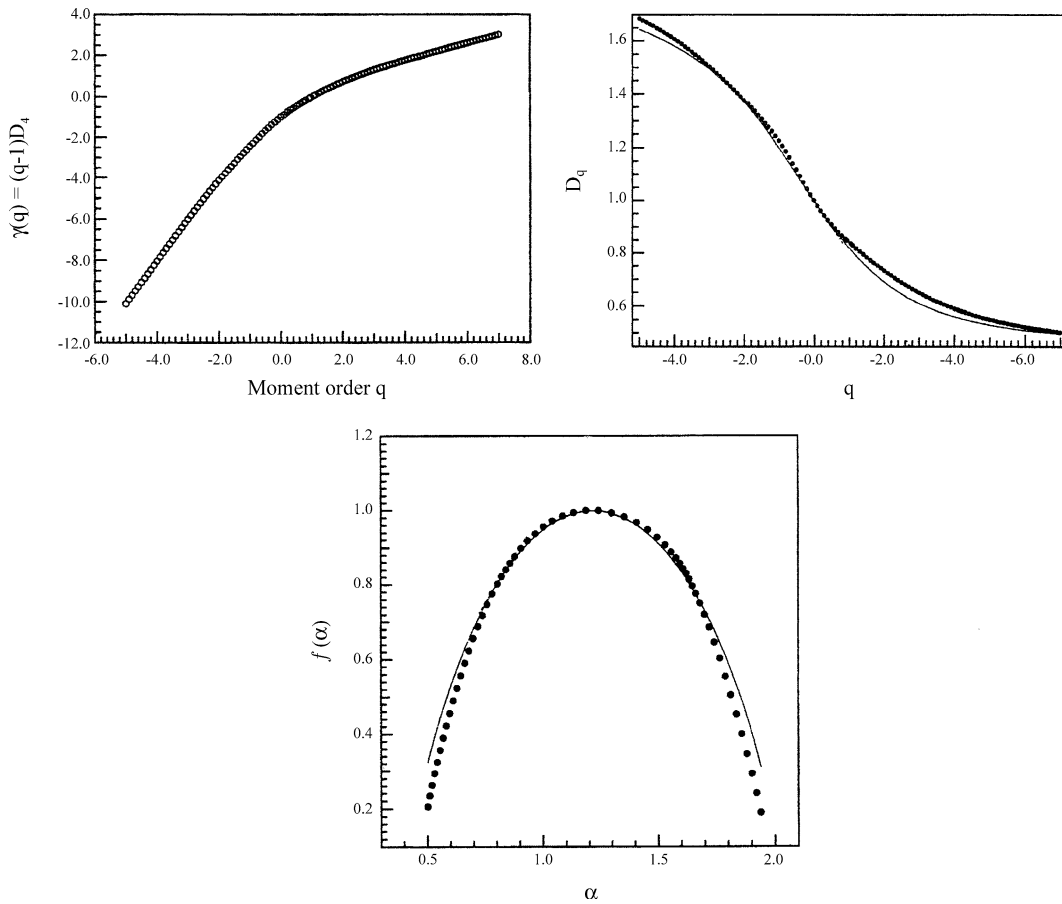
$$\gamma(q) = D_q(1 - q) = [q\alpha(q) - f(\alpha(q))] \quad (37)$$

Differentiating (37) with respect to  $q$  and using the first expression of (35), we obtain

$$\alpha(q) = \gamma'(q) \equiv d\gamma/dq \quad (38)$$

Thus, given  $\gamma(q)$ , we may determine  $\alpha(q)$  and then,  $f(\alpha)$ . Such a singularity spectrum for the AE index considered by Consolini et al. [41] is given in Fig. 15 (bottom panel). Because  $\alpha(q)$  is intimately related to  $\gamma(q)$  (and therefore





**Space Plasmas, Dynamical Complexity in, Figure 15**

**Top left:** Partition function exponent  $\gamma(q)$  for a time series of the AE (auroral electrojet) index. **Top right:** Generalized dimension  $D_q$  for the same AE time series. Solid line is the best fit using the P-model. **Bottom:** Singularity spectrum  $f(\alpha)$  for the same AE time series obtained from the Legendre transform. Adapted from [41] with thanks to the American Physical Society

also to the partition function  $\Gamma(q)$ ,  $f(\alpha)$  sorts out the spectrum of singularities of the intermittent fluctuations, that may be mapped to the multifractal dimensions  $D_q$  as defined by (32).

In the above, we have introduced a sequence of algebraic manipulations mainly for the purpose of connecting three different popular approaches of identifying the singularity nature of the multifractal characteristics of a time (or spatial) series based on the partition function formulation. We must emphasize that these are merely different approaches of viewing fracture characteristics of intermittent fluctuations with no additional physics involved, except perhaps with the following useful thermodynamic analogy. That is, the relationship between  $(\gamma(q), q)$  – or equivalently  $(D_q, q)$  – and  $(f(\alpha), \alpha)$  is formally a Legendre transform. Thus, one may easily associate the relationships among such measures with the standard notion of Leg-

endre transforms of classical thermodynamics [96]. Such a connection may become useful in identifying analogies between singularity measures of multifractals with multicriticality measures in thermodynamics.

It is interesting to note that in multifractal analyses of space plasma turbulence in regions where intermittency are expected, similar functional dependence of  $\gamma(q)$  on  $q$  as above is also generally obtained [12,13,14,15,18,34,87,134,139]. Although, we must recall that for turbulence in the plasma sheet or the auroral zone and even sometimes in the solar wind, the assumption of the validity of the Taylor hypothesis must be applied with caution.

In the previous section, it was demonstrated that intermittency characteristics of a fluctuating series could be brought out using the technique of wavelet transforms. Certain quantities, Local Intermittency Measure LIM or Normalized Power and Flatness, were introduced using

the wavelet transforms to gauge the intermittency effects. Since the complete set of coefficients of the wavelet transforms essentially represents the full fluctuating event, it is reasonable to inquire if it can be used to advantage to describe the full characteristics of intermittency of a fluctuating series. The answer is a definite *yes*. For example, wavelet transforms have been used to study multifractal characteristics of time series related to intermittent turbulence in space plasmas by Yordonova et al. [139] for the magnetospheric cusp and Vörös et al. [131] for the plasma sheet.

### Rank-Ordered Multifractal Analysis (ROMA)

In the above, intermittent fluctuations were analyzed using the structure function and/or partition function methods based on the statistics of the full set of fluctuations. Since most of the observed or simulated intermittent fluctuations are dominated by fluctuations with small amplitudes, the subdominant fractal characteristics of the minority fluctuations – generally of larger amplitudes – are easily masked by those characterized by the dominant population. It therefore appears useful to search for a procedure that explores the fractal nature of the subdominant fluctuations by first appropriately isolating out the minority populations and then perform statistical investigations for each of the isolated populations.

Phrasing it differently, for intermittent fluctuations exhibiting multifractal characteristics, we visualize that these fluctuations are composed of many types, each type being characterized by a particular fractal dimension. The questions to ask then are (i) what are the different fractal dimensions? and (ii) how are the various types or classes of fluctuations distributed within the turbulent medium? We answer these questions quantitatively by constructing a “rank-ordered multifractal spectrum”.

Using the example of a 2D MHD simulation that was discussed in Sect. “Non-Gaussian Probability Distribution Functions”, we demonstrate below how this idea may be achieved with a rank-order method that subdivides the fluctuations into groupings based on the range of the scaled-sizes of the fluctuations [28]. In the simulation, ideal compressible MHD equations expressed in conservative forms are solved numerically with  $1024 \times 1024$  grid points in a doubly periodic  $(x, z)$  domain of length  $2\pi$  in both directions using the WENO code [64] so that the total mass, energy, magnetic fluxes and momenta are conserved. The initial condition consists of random magnetic field and velocity with a constant total pressure for a high beta plasma. As discussed previously, after sufficient elapsed time, the system evolves into a set of randomly in-

teracting multiscale coherent structures exhibiting classical aspects of intermittent fluctuations.

We recall that the PDFs  $P(\delta B^2, \Delta)$  of the square of the magnitude of the magnetic field  $\delta B^2$  for such intermittent fluctuations were non-Gaussian and became more and more heavy-tailed at smaller and smaller scales (Fig. 8). An attempt to collapse the unscaled PDFs according to the one-parameter scaling formula (12) re-displayed below as (39):

$$P(\delta B^2, \delta) \delta^s = P_s(\delta B^2 / \delta^s) \quad (39)$$

indicated approximate scaling with an estimated scaling exponent of  $s = 0.335$  (Fig. 9). The scaling formula (39) may be interpreted from the concept of fractals. We begin the interpretation by assuming that  $\delta B^2$  and  $P$  vary with  $\delta$  as power laws:  $\delta B^2 / \delta^a = I$  and  $P / \delta^b = J$ , where  $(a, b)$  are the power-law (fractal) exponents and  $(I, J)$  are constants – i. e. invariants with respect to the scale  $\delta$ . If the form of  $P(\delta B^2, \delta)$  is invariant as the scale changes, then it has been shown that a functional relation exists between the two invariants,  $(I, J)$  [29]. Imposing the normalization condition for the PDFs, we obtain the one-parameter scaling form (39) with  $s = a = -b$ .

Such a phenomenon is self-similar and the PDFs collapse onto one scaled PDF  $P_s(Y)$  where  $Y = \delta B^2 / \delta^s$  is a scale invariant. In such a case, the structure function exponents for the PDFs will satisfy the monofractal property of  $\zeta_q = sq$ . The scaling exponent  $s$  may be interpreted as a single fractal measure that characterizes the fluctuations of all scales through the scaling relation (39). If the PDFs are self-similar Gaussian distributions for all scales, representing random diffusion, the scaling exponent  $s$  is equal to 0.5. For other monofractal distributions, the scaling exponent may take on any positive real value.

For the current example, structure function calculations based on the full set of simulated fluctuations showed a nonlinear relation between the exponents and the moment order, indicating the phenomenon is multifractal in nature [27]. However, the physical interpretation of the multifractal nature is not easily deciphered by merely examining the curvature of the deviation from linearity.

We shall therefore attempt to perform statistical analyzes individually for subsets of the fluctuations that characterize the various fractal behaviors within the full multifractal set. Such grouping of fluctuations must depend somehow on the sizes of the fluctuations. However, the groupings cannot depend merely on the raw values of the sizes of the fluctuations because the ranges will be different for different scales. Thus, we proceed to rank-order the sizes of the fluctuations based on the ranges of the scaled variable  $Y = |\delta B^2| / \Delta^s$  where  $s$  is the scaling exponent for

each grouping defined in (39). (Absolute values of the fluctuations are used to take advantage of the symmetry property of the PDFs and for the purpose of better statistical convergence.) For each chosen range we shall assume that the fluctuations of all scales will exhibit monofractal behavior satisfying the form of (39) and be characterized by a scaling exponent  $s$ . The question is then how can this procedure be accomplished.

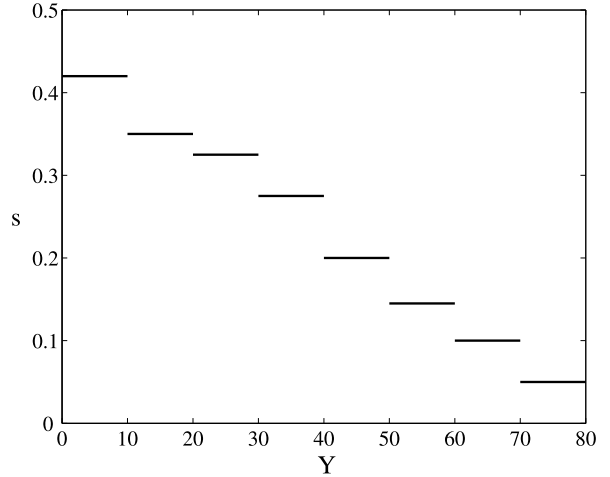
We begin by considering a differential range of scales  $dY$  in the vicinity of some scaled size  $Y = |\delta B^2|/\Delta^s$ . Fluctuations whose sizes fall within such a differential range will probably exhibit monofractal behavior characterized by some scaling exponent  $s$  such that the differential structure function  $dS_q$  will vary with the scale as  $\Delta^{sq}$ , i. e.,

$$dS_q \triangleq (|\delta B^2|)^q P(|\delta B^2|, \Delta) d(|\delta B^2|) = \Delta^{sq} Y^q P_s(Y) dY \quad (40)$$

Given a set of PDFs  $P(|\delta B^2|, \Delta)$ , the corresponding multifractal spectrum  $s(Y)$  may be obtained approximately (if the ansatz is valid) by integrating the functional differential expression (40) over small contiguous ranges of  $\Delta Y$  with the assumption that within each incremental range the scaling exponent  $s$  is essentially a constant [28]. Thus, for a generic range of  $\Delta Y$  within  $(Y_1, Y_2)$ , we form such a range-limited structure function as follows:

$$\begin{aligned} \Delta S_q(|\delta B^2|, \Delta) &= \int_{a_1}^{a_2} |\delta B^2|^q P(|\delta B^2|, \Delta) d|\delta B^2| \\ &\simeq \Delta^{sq} \int_{Y_1}^{Y_2} Y^q P_s(Y) dY \end{aligned} \quad (41)$$

where  $a_1 = Y_1 \Delta^s$  and  $a_2 = Y_2 \Delta^s$ . We then search the value of  $s$  such that the scaling property of the range-limited structure function that varies with  $s$  is  $\Delta S_q(s) \sim \Delta^{sq}$ . If such a value of  $s$  exists, then we have found one region of the multifractal spectrum of the fluctuations such that the PDFs in the range of  $\Delta Y$  collapses onto one scaled PDF. Performing this procedure for all contiguous ranges of  $\Delta Y$  produces the rank-ordered multifractal spectrum  $s(Y)$  that we are looking for. The determined value of  $s$  for each grouping should be unaffected by the statistics of other subsets of fluctuations that are not within the chosen range  $\Delta Y$  and therefore should be quantitatively quite accurate. The physical meaning of this spectrum is that the PDFs for all time lags collapse onto one master multifractal scaled PDF. The spectrum is implicit since  $Y$  is a defined as a function of  $s$ .



Space Plasmas, Dynamical Complexity in, Figure 16

Rank-ordered multifractal spectrum,  $s(Y)$ .  $Y = |\delta B^2|/\Delta^s$ ;  $|\delta B^2|$  in units of bin size and  $\Delta$  in grid spacing. The spectrum is calculated for 8 contiguous ranges of  $\Delta Y$ . Reproduced from [28] with thanks to the American Physical Society

Interestingly, for this particular example, there exists one and only one value of  $s$  in each range of  $\Delta Y$ , that satisfies the above constraint. Unlike the structure functions defined for the full range of fluctuations, the range-limited structure functions based on (41) exists also for negative real values of  $q$  (except for the range including  $Y = 0$ ). Figure 16 displays the calculated rank-ordered spectrum  $s(Y)$  based on eight contiguous ranges of  $\Delta Y$  [28]. It is noted that the spectrum has values of  $s$  ranging between 0.5 and 0.0. The spectrum can be refined by choosing more range intervals with smaller range sizes of  $\Delta Y$ , although in practice this procedure is limited by the availability of simulated data points. At  $Y = 0$ , the scaling exponent  $s$  appears to approach the self-similar Gaussian value of 0.5 representing random diffusion. As the value of the scaled fluctuation size  $Y$  increases the scaling exponent decreases accordingly indicating the fluctuations are becoming less and less space filling. At very large value of  $Y$ , the scaling exponent  $s$  seems to asymptotically approach the value of zero indicating the fluctuations have become extremely sparse.

Such an implicit multifractal spectrum has several advantages over the results obtainable using the conventional structure function calculations. Firstly, the utility of the spectrum is to fully collapse the unscaled PDFs. Secondly, the physical interpretation is clear. It indicates how space-filling (in terms of the value of  $s$ ) are the scaled fluctuations once the value of  $Y$  is given. Thirdly, the determination of the values of the fractal nature of the grouped fluctuations

is not affected by the statistics of other fluctuations that do not exhibit the same fractal characteristics. Fourthly, it provides a natural connection between the one-parameter scaling idea (39) and the multifractal behavior of intermittency.

## Invariant Scaling

### Forced and/or Self-Organized Criticality (FSOC)

In the above sections, we provided some convincing arguments as well as numerical and observational evidences indicating that space plasma turbulence is generally in a state of dynamic “topological complexity”. By “complex” topological states we mean magnetic topologies that are not immediately deducible from the basic (e. g., MHD) equations. We also discovered that the structure/partition functions for turbulence generally scale as power laws at least at small scales. Expressed another way, we found that there were power-law scale invariants with respect to coarse-graining at small scales where intermittency prevails. Below, we shall briefly address the salient features of such “topological phase transitions” as well as the associated scale invariant properties in turbulent plasmas. (We note here that the concept of phase transitions has also been employed to address the phenomenon of complexity in space plasmas by Sitnov et al., [106] using the singular spectrum method.)

For nonlinear stochastic systems exhibiting complexity, the correlations among the fluctuations of the random dynamical fields (electric, magnetic and velocity fields, etc.) are generally very long-ranged and there exist many correlation scales (as exhibited by the various outsized plasma coherent structures). The dynamics of such systems are notoriously difficult to handle either analytically or numerically. On the other hand, since the correlations are extremely long-ranged, it is reasonable to expect that the system will exhibit some sort of invariance property under coarse-graining scale transformations. A powerful analytical technique that exploits this invariance property is the method of the dynamic renormalization group (DRG) which we shall briefly describe below. The technique is a generalization of the static renormalization group introduced by Wilson in 1972 [135] for equilibrium phase transitions. Dynamic renormalization group was originally developed as a perturbation theory by Halperin et al. [53] and first applied to hydrodynamic problems by Forster et al. [49]. A closed form exact theory of DRG was developed by Chang et al. [30]. And a review of the basic elements of this theory may be found in Sect. “Non-Gaussian Probability Distribution Functions” of a Physics Report by Chang et al. [31].

No matter how complex the stochastic behavior is for the turbulent system, we may conjecture that the system is somehow characterized by a set of dynamically evolving parameters  $\{P_n\}$ . We now transform the stochastic system by coarse-graining and ask how these parameters transform accordingly (subject to the rules stipulated by the underlying dynamical equations, e. g., those for the interacting flux tubes). Symbolically, we write

$$\partial\{P_n\}/\partial\ell = R\{P_n\} \quad (42)$$

where  $R$  represents the coarse-graining (renormalization-group) transformation operator and  $\ell$  is the coarse-graining parameter. The operator is generally very complicated for even relatively simple stochastic systems. Instead of actually delving into the mathematical formalism of DRG, we shall merely explore some of the basic concepts related to such ideas.

**Forced and/or Self-organized Criticality** Generally, if one starts at some initial state  $\{P_n(0)\}$  and proceeds with the coarse-graining procedure following the prescription dictated by (42), the result would be a trajectory in the phase space characterized by the parameters. In analogy to nonlinear dynamics, we expect that there may be fixed points (singular points) in the phase space flow field, at which  $d\{P_n\}/d\ell = 0$ . At such a point, e. g.,  $\{P_n^*\}$ , the correlation length should not be changing. However, the coarse-graining transformation requires that all length scales to change with  $\ell$ . Therefore, to satisfy both requirements, the correlation length must be either infinite or zero at the fixed point. When it is at infinity, the state of the dynamical system would then be at a state of dynamic criticality, analogous to the state of criticality in equilibrium phase transitions.

Generally, a dynamical system will not be near a parametric state such as  $\{P_n^*\}$ . But, if the coarse-graining procedure (i. e., the renormalization group trajectory) leads it eventually to a point in the parametric space close to a fixed point such as  $\{P_n^*\}$ , then in a coarse-grained sense the stochastic system would be close to a state of dynamic criticality. We may then approximate the coarse-graining transformation (42) by linearizing the operator  $R$ . The result is a set of coupled linear differential equations characterizing the variations of the deviations of the parameters  $\{P_n\}$  from  $\{P_n^*\}$ .

$$d\{Q_n\}/d\ell = R_L\{Q_n\} \quad (43)$$

where  $Q_n = P_n - P_n^*$  and  $R_L$  is the linearized renormalization-group matrix operator.

We assume that expression (43) may be expressed in terms of a set of eigenvalue equations:

$$d\mu_i/dl = \lambda_i \mu_i \quad (44)$$

where the eigenvectors  $\mu_i$ 's are linear combinations of the parameters  $Q_n$  and the  $\lambda_i$ 's are the corresponding eigenvalues. We shall assume that the eigenvectors are complete and the eigenvalues are all real. (Nondiagonalizable operator  $R_L$  may render the eigenvectors incomplete and lead to logarithmic corrections to the discussion below, which will not be considered here).

The solutions of (44) are simply  $\mu_i = \mu_i^0 \exp \lambda_i \ell$  where  $\{\mu_i^0\}$  represents some arbitrary state that is close to  $\{P_n^*\}$ . From this, it can be verified immediately that

$$\mu_i / \mu_j^{\lambda_i/\lambda_j} = \text{constants} \quad (45)$$

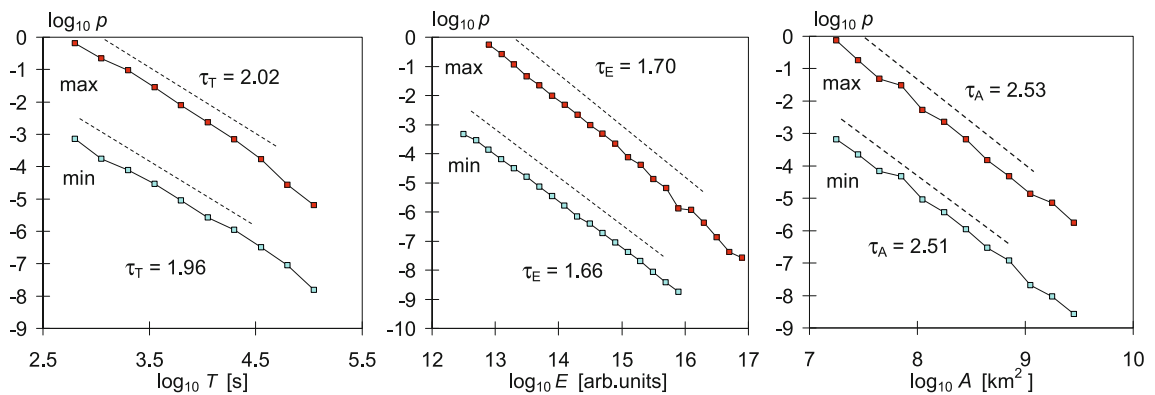
i. e., the  $\mu_i$ 's form power law invariants which are independent of the coarse-graining scale transformation  $\ell$ .

Expressing this result in words: The mathematical consequence of the linear approximation is that, close to dynamic criticality, certain linear combinations of the perturbations of the parameters that characterize the stochastic phenomenon will correlate with each other in the form of power laws. These include the  $(k, \omega)$ , i. e. mode number and frequency, spectra of the correlations of the various fluctuations of the dynamic field variables in space and time, etc. The fact that the dynamical (stochastic) system must initiate from some state so that the coarse-graining trajectory will carry it to the close proximity of a fixed point such as  $\{P_n^*\}$  means that there may need some initial tuning of the state for the system to have such behavior. For this reason, we shall refer to such dynamical systems

as systems near Forced and/or Self-organized Criticality (FSOC) [22]. The phenomenon of FSOC is spatiotemporal and the invariant scaling idea permeates throughout the time-space domain. In addition, the scaling between time and space generally is not isotropic.

Such power law behavior has been detected in probability distributions related to the solar flares [74,126], Fig. 17, in the AE burst occurrences as a function of the AE burst strength [37], Fig. 18, in the global auroral UVI imagery of the statistics of size and energy dissipated by the magnetospheric system [81], in the probability distributions of spatiotemporal magnetospheric disturbances as seen in the UVI images of the nighttime ionosphere [124], in the probability distributions of durations of Bursty Bulk Flows [2], and in the magnetic field fluctuations of the Earth's magnetotail [61,93]. Similar behavior has also been obtained in numerical simulations for the distributions for solar flares [45], for dynamical models for the current sheet [66,110,122] and other applications. This invariant behavior had been called self-organized criticality (SOC) [3], i. e., tuning is generally not necessary.

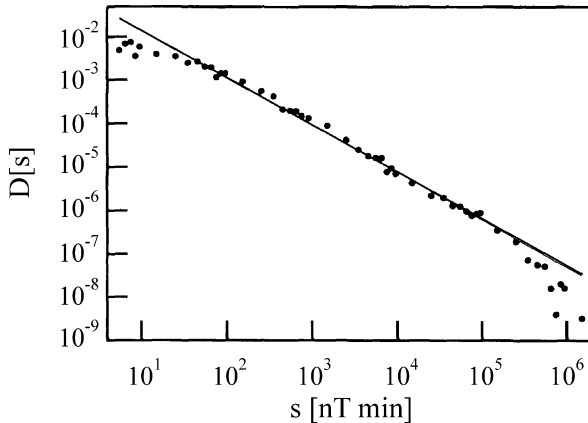
An eigenvector (eigen-parameter)  $\mu_i$  is "irrelevant" if the corresponding eigenvalue  $\lambda_i$  is less than zero because in the limit of large  $\ell$  the value of  $\mu_i$  will become negligible. The reverse is true for positive  $\lambda_i$ 's. We shall not consider the special case of  $\lambda_i = 0$  which introduces singularities of fractional powers of logarithms. If there are only a small number of relevant eigenvalues with  $\lambda_i > 0$ , the corresponding FSOC system has only a small number of "relevant parameters" that are needed to characterize the stochastic state, i. e., the system may appear to have low dimensionality in parameter space near criticality [22].



Space Plasmas, Dynamical Complexity in, Figure 17

Flare probability distributions of avalanche lifetime (left), emission flux (middle) and peak area (right) for solar max (high activity) and min (low activity) obtained for emissions detected by the EIT (extreme ultraviolet imaging telescope) aboard the SOHO spacecraft in the 195 Å wavelength above a threshold of 40% of the average emission. The min distributions are shifted downward for comparison. Adapted from [126] with thanks to the American Physical Society





Space Plasmas, Dynamical Complexity in, Figure 18

Distribution  $D(s)$  of the burst size  $s = \int_{\Omega} [AE(t) - L]dt$  of an AE index time series, where the integration is performed over the time interval  $\Omega$  when the quiet level of  $L = (45 \pm 15)nT$  is exceeded. The solid line is a power-law best fit. Reprinted from [37] with permission from the Italian Society of Physics

### From Criticality to Intermittency

We now ask a deeper question: How does the invariant behavior of the structure or partition functions, which changes with moment order for intermittent turbulence amalgamate (reconcile) with the concept of criticality (SOC or FSOC) outlined above?

For systems near criticality the correlation functions are generally related to the correlation time and space as power laws. Since the structure functions are related to the spatial correlation functions [7], it is therefore not surprising that for systems near SOC/FSOC, the structure function exponents for intermittent turbulence are power law scale-invariants in coarse-graining scale transformations.

For a stochastic process that is monofractal, the structure function exponents vary with the moment order linearly. Thus a single invariant, such as  $I_1 = S_1/\Delta^{\zeta_1}$  would suffice to characterize the behavior of fully developed turbulence that is self-similar. For example, it was demonstrated in Subsect. “Rank-Ordered Multifractal Analysis (ROMA)”, that if the probability distributions functions (PDFs) satisfy the one-parameter scaling law of (39), the structure function exponents  $\zeta_q = sq$  where  $s$  is the scaling exponent, and the PDFs collapse onto one master scaling function  $P_s(\delta B^2/\Delta^s)$ . This is approximately satisfied by the PDFs for the numerical simulation results discussed in Sect. “Non-Gaussian Probability Distribution Functions” [33] (Fig. 9 and re-displayed here as Fig. 19, middle panel). Similar results have been found by Hnat et al. [56] for their study of the PDFs of solar wind turbulence as seen by the WIND spacecraft and Tam et al. [112] for

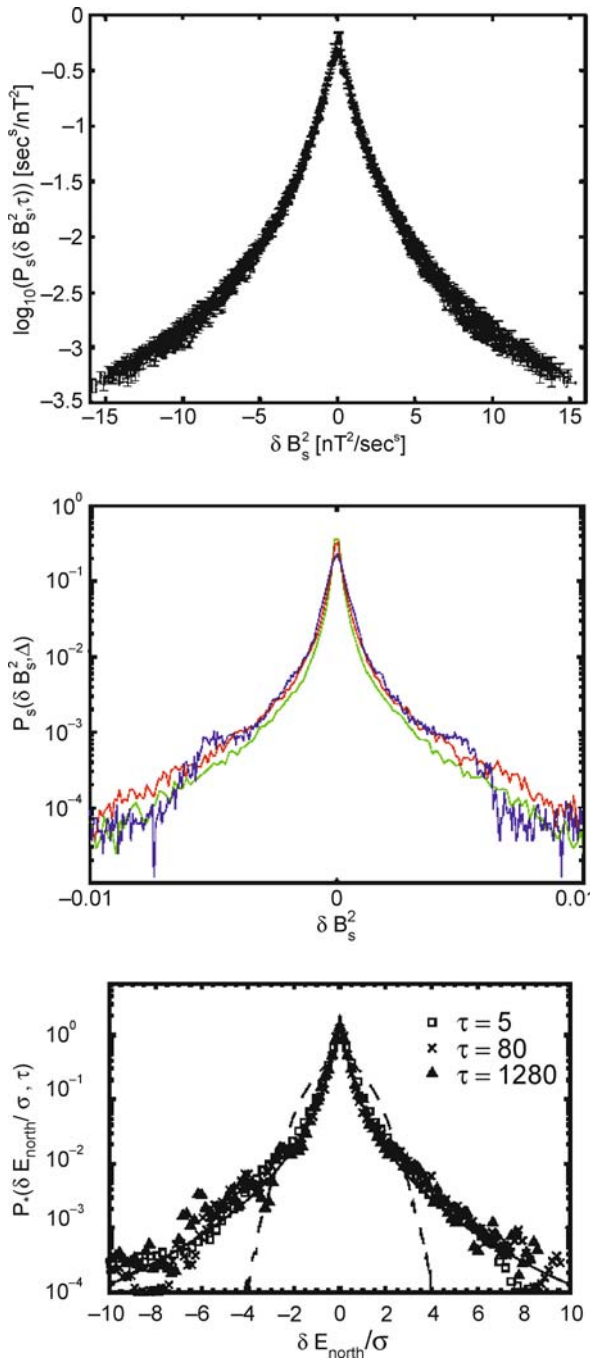
the electric field intensity fluctuations in the auroral zone as detected by the SIERRA rocket experiment (Fig. 19, top and bottom panels). However, it is to be noted that in all three scaling plots the deviations from the mean become larger for larger fluctuations, which are indications that multifractal characteristics will still need to be addressed. This was demonstrated in terms of the rank-ordered multifractal spectrum as discussed in Subsect. “Rank-Ordered Multifractal Analysis (ROMA)”.

We note that generally – e.g., Fig. 8 – the portions of the PDFs for small fluctuations for most of the observed and simulated intermittent turbulence have shapes very close to that of a Gaussian. This is the reason why the calculated exponent  $s$  for small scaled fluctuations is close to the value of 0.5 (characteristic of self-similar random diffusion). In the numerical example considered in Subsect. “Plasma Resonances and Coherent Structures”,  $s \approx 0.42$  for the range of  $\Delta Y$  between (0, 10). And the behavior of the fluctuations within this range was shown to be essentially monofractal. This would have given the value of the first structure function exponent  $\zeta_1 = 0.42$  and the related power spectrum index would therefore be approximately  $1 + \zeta_2 = 1 + 2\zeta_1 \approx 1.84$  (Frisch, 1995). If no attempt were made in rank-ordering and the value of  $s$  were determined for the full set of fluctuations by looking for power-law behavior of  $P(0, \Delta) \sim \Delta^s$ , the best value was found to be approximately 0.335. Assuming approximate monofractal scaling, this number would give a power spectrum index of approximately 1.67. Both numbers are not very different from that obtainable using classical arguments for hydrodynamic or MHD turbulence.

Most observational or numerical “critical exponents” of SOC are obtained either using first order box counting or from calculations of density, optical, and other field intensities such as electric, magnetic, velocity, and energy fields, which generally involve statistical averages over the zeroth, first, or second (i.e. low) order moments of their related PDFs and therefore each represents an approximate monofractal behavior of some sort of weighted average of the fractal values for small scaled fluctuations.

In fact, if the experimental or numerical evaluations are based on rough estimates, the values of the exponents will generally take on essentially the Gaussian – i.e., mean-field – values. And such observed “critical exponents” would satisfy the standard scaling relations. We suggest that this is the reason why recent observations [126] and numerical simulations [69] seem to indicate the coexistence of the phenomenon of SOC and multifractal signatures for intermittent turbulence. See, e.g. Figs. 17 and 20.

Thus, based on the example of rank-order procedure of the multifractal behavior of intermittent fluctuations



described in Subsect. “[Rank-Ordered Multifractal Analysis \(ROMA\)](#)”, we suggest that power-law relations of SOC can be made more precise if the exponents are determined more meticulously using rank-ordered fluctuations. For each rank-order, particularly for the rank-order of the smallest scaled fluctuations, one can determine a corre-

#### ◀ Space Plasmas, Dynamical Complexity in, Figure 19

*Top:* Scaled PDF based on the one-parameter scaling form of (39) with  $\zeta_1 = s \approx 0.42$  for the fluctuations of the magnetic field energy density as seen by the WIND spacecraft. The data shown correspond to  $\tau$  between 46 seconds and 26 hours. Adapted from [56] with thanks to the American Geophysical Union. *Middle:* Scaled PDF for magnetic field energy density fluctuations based on the one-parameter scaling form of (39) with  $\zeta_1 = s \approx 0.335$  from a 2D MHD simulation for homogeneous turbulence for  $\Delta = 2$  (green), 8 (red), 32 (blue) units of grid spacing. Adapted from [33] with thanks to the American Institute of Physics. *Bottom panel:* Normalized PDF for fluctuations of transverse electric field intensity  $E_{\text{north}}/\sigma$  at  $\tau = 5, 80, 1280$  ms, where  $\sigma = \tau^{\zeta_1}$  ( $\zeta_1 \approx 0.8$ ) is the root mean square of the fluctuations. The solid line corresponds to the normalized Castaing distribution with  $\lambda = 1.02$ . The results of the analyzes for Gaussian fluctuations (dashed) are shown for comparison. Reprinted from [112] with thanks to the American Geophysical Union

sponding set of more accurately determined “critical exponents” that satisfy the classical exponent relations.

In summary, the conventional concepts of SOC or FSOC as tested by the lower order spatiotemporal statistical averages would generally be satisfied for intermittent turbulence. However, the ideas of invariant scaling and scaling relations will need to be generalized to include the “crossover effects” for large scaled fluctuations.

#### Dynamical Modeling – The Lu-Klimas Magnetic Field Reversal Model

By utilizing results obtained from direct numerical simulations, investigators were able to confirm a number of the theoretical predictions of intermittent turbulence and complexity phenomena in space plasmas that could arise from the process of merging and interactions of plasma coherent structures. The system sizes of direct numerical simulations, however, are generally limited by the capabilities of the present day computers. To circumvent this limitation, an alternative approach to study the intermittency in space plasma turbulence numerically would be to mimic the coarse-grained dissipation and the development of complexity by means of dynamical modeling which, in turn, may lead to a better theoretical understanding of the fundamental underlying stochastic processes.

While there exists a number of dynamical models that mimic avalanche processes exhibiting scaling and some also intermittency, only a few have attempted to incorporate the basic physics of plasmas in their formulations. From these, we select one simple model, the Lu-Klimas fast-field annihilation model of magnetic field reversal, for further discussion and comments. The model has the basic

ingredient of a diffusive equation [66]:

$$\frac{\partial B_x}{\partial t} = \frac{\partial}{\partial z} \left[ D(z, t) \frac{\partial}{\partial z} \right] B_x + S(z, t). \quad (46)$$

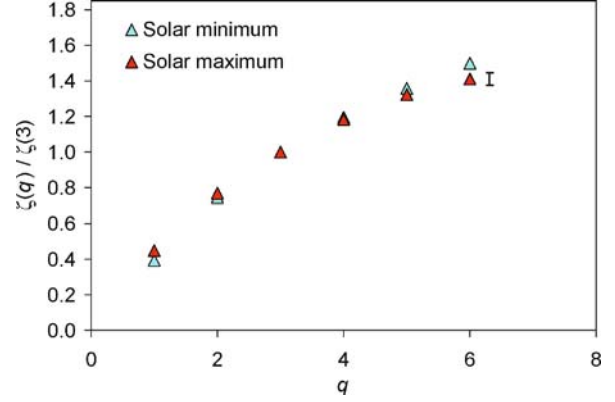
It may be obtained through the drastic reduction of the resistive MHD system in an idealized one-dimensional limit. In (46),  $B_x(z, t)$  is the magnetic flux and  $S(z, t)$  is a source term which could be interpreted as that characterizes the convection of the magnetic flux, although in discussions below it will be arbitrarily prescribed and assumed to be independent of time. The diffusion coefficient  $D(z, t)$  is linearly proportional to an anomalous resistivity with a hysteresis characteristic as stipulated below:

$$\frac{\partial D}{\partial t} + \frac{D}{\tau} = \frac{Q(|\partial B_x / \partial z|)}{\tau} \quad (47)$$

where  $Q = D_{\min}$  (low state) for  $|\partial B_x / \partial z| < k$ , and  $Q = D_{\max} \gg D_{\min}$  (high state) for  $|\partial B_x / \partial z| > \beta k$  with  $\beta < 1$ . Thus,  $Q$  is double-valued and dependent upon the history for  $\beta k < |\partial B_x / \partial z| < k$ . The value  $k$  may be thought of as a stability threshold which Klimas et al. [66] attributed to some sort of current driven instability (e.g., Lui et al. [79]). The value of  $Q$  remains in the low state ( $= D_{\min}$ ) until the threshold is reached and switches to the high state ( $= D_{\max}$ ). It stays at the high state until  $|\partial B_x / \partial z| < \beta k$ . This anomalous diffusive model is essentially that originally suggested by Lu [73], which was motivated by his studies related to the phenomena of self-organized criticality [3]; in particular, those related to the probability distributions of solar flares [74]. Unlike the Klimas' model, the source term in Lu's original formulation was assumed to be random and the boundary conditions were chosen to allow transport through the boundaries.

Klimas et al. [66] and Uritsky et al. [123] have studied their model numerically on a spatial interval  $-L \leq z \leq L$  subject to the boundary conditions  $\partial B_x / \partial z = 0$  at  $z = \pm L$ , i.e., no transport of  $B_x$  through the boundaries. They considered antisymmetric solutions with a field reversal at  $z = 0$  and a time-independent source term  $S(z, t) = S_0 \sin(\pi z / 2L)$  which may be considered as a steady convection of  $B_x$  into the field reversal region. Setting  $\tau = 1$  as the fundamental time scale and  $L = 20$ , they found that there exists a range of choices of parameters for which the dynamical model evolves in a sequence of quasi-periodic loading ( $dE/dt > 0$ )-unloading ( $dE/dt < 0$ ) cycles where the total field energy  $E(t)$  is defined by:

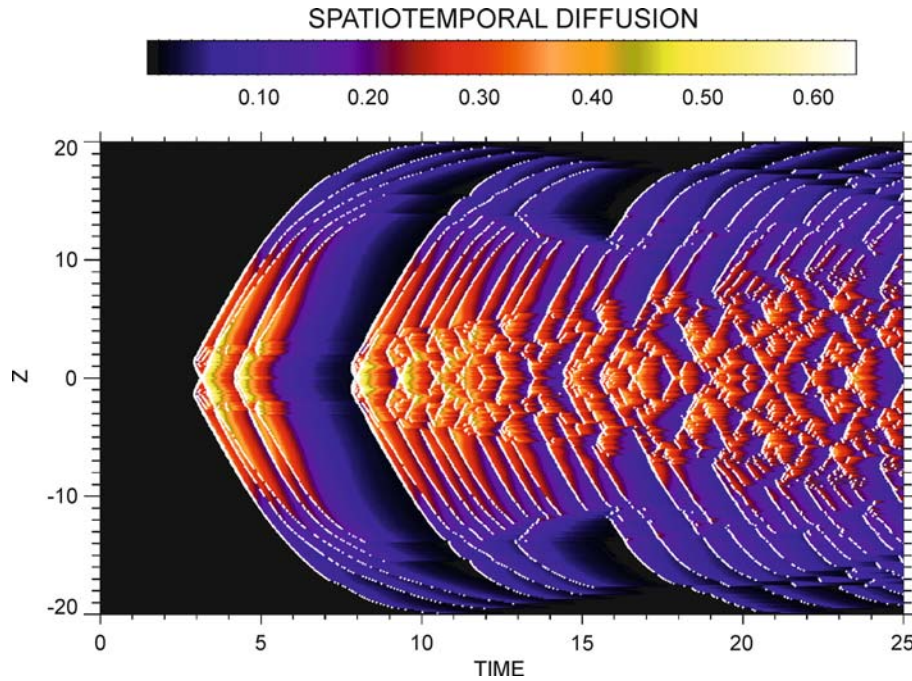
$$E(t) \equiv \int_{-L}^L dz B_x^2(z, t). \quad (48)$$



Space Plasmas, Dynamical Complexity in, Figure 20

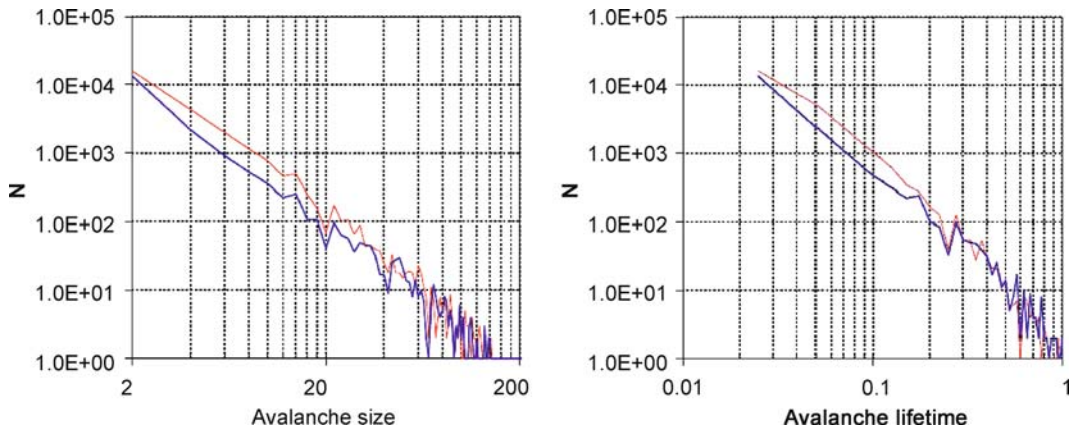
Relative structure function exponents with respect to the third order exponent using ESS for flare statistics at solar maximum (high activity) and solar minimum (low activity) based on emissions detected by the EIT (extreme ultraviolet imaging telescope) aboard the SOHO spacecraft in the 195 Å wavelength. Adapted from [126] with thanks to the American Physical Society

During the loading intervals, when the system is quiet, the slow increase of  $E(t)$  is the sole evolutionary feature of the system. As loading continues, the field gradient  $|dB_x/dz|$ , i.e. the current density, reaches the critical level for  $Q$  somewhere in the field distribution (usually at a pair of points symmetrically placed near  $z = 0$ ). Subsequently a complicated process of unloading ensues, which is best understood in terms of the spatiotemporal evolution of the diffusion coefficient  $D(z, t)$  as depicted in Fig. 21 during the initial stage of development of the unloading process, where the color scale of the figure indicates the intensity of the diffusion coefficient. In the figure, a white dot is placed at each point where the critical level of  $|dB_x/dz|$  – or current density – for  $Q$  is reached and  $Q(z, t) = D_{\max}$ . As the instabilities spread and propagate toward both the center and the edges of the field distribution, more current sheets (i.e., sheet regions with strong current density) are excited and their topology becomes increasingly complex. The current sheets generate diffusion and the diffusion coefficient  $D(z, t)$  is at a local maximum just after the passage of a current sheet at any given position. This strong localized diffusion enhances the transport of the magnetic flux toward  $z = 0$  where fluxes of opposite signs meet and annihilate. This process of generation of complicated current sheet topology, transport of magnetic flux and annihilation adjusts self-consistently with the imposed source function such that there is a dynamically pseudo-stationary state. And, the loading-unloading process repeat quasi-periodically after an initial transient.



Space Plasmas, Dynamical Complexity in, Figure 21

Spatiotemporal evolution of the diffusion coefficient  $D(z, t)$  for the Lu-Klimas model during the initial stage of development of an unloading process. The color scale of the figure indicates the intensity of the diffusion coefficient. A white dot is placed at each point where the critical level for  $Q$  is reached and  $Q(z, t) = D_{\max}$ . Reprinted from [66] with thanks to the American Geophysical Union



Space Plasmas, Dynamical Complexity in, Figure 22

Avalanche size (left) and lifetime (right) distributions for the Lu-Klimas model for early (blue) and late (red) intervals during a single unloading event. Reprinted from [123] with permission from Elsevier

The above phenomenon of continued sporadic generation of unstable current sheets and the accompanying broadening of the diffusive layer during the unloading process is reminiscent of the resonance broadening phenomenon of the “apparent reconnection” process in the 2D MHD simulation result for the sheared magnetic field geometry discussed earlier. Klimas et al. [67,68,69],

Uritsky et al. [125] and Uritsky and Klimas [122] have recently extended their dynamical modeling effort based on the anomalous resistivity idea with hysteresis to the 2D geometry and the similarity between their calculated results of sporadic localized reconnections and the direct 2D MHD simulation results of resonance-broadening and coarse-grained dissipation – as well as the ensuing mag-



netic reconfiguration process – due to the sporadic, localized merging of coherent structures is even more noticeable, even though the boundary and initial conditions as well as the presumed instabilities of both sets of calculations were quite different. The main point to note is that the phenomenon of coarse-grained dissipation in the results of direct numerical simulation is replaced in the dynamical model by the assumed anomalous resistivity with hysteresis. Thus, other than the arbitrariness of the dissipation mechanism, such type of dynamical modeling has almost achieved its purported aim; i. e., using a much simplified dissipation model, large scale simulations of complexity of space plasmas may perhaps be achieved with relatively small scale computers.

However, the existing dynamical models may be too crude to lead to quantitatively comparable results of actual observations. For example, in the Lu–Klimas model, the critical threshold was arbitrary and kept at a constant value. In addition the phenomenon of hysteresis was again quite arbitrary and the double-valuedness of  $Q$  was also triggered at arbitrary levels. In the coarse-grained dissipation scenario acquired from, for example, direct numerical simulations, there was actually no specific threshold for triggering the “fluctuation induced nonlinear instability” and the dissipation during “apparent reconnection” was due to the combination of wave (e. g., Alfvén wave) propagation and resonance broadening. The hope is that perhaps even with some of such arbitrary ad hoc assumptions in the present and/or future modeling of local dissipation, the hypothesized model(s) will still lead to some of the global features such as the scale-invariant properties that will fall in the same universality class as the actual physical phenomena.

We now proceed to discuss the invariant properties of the simulated result of the one-dimensional magnetic field reversal and annihilation model. There are at least two sets of statistics that may shed light on the scale free invariant properties of SOC for the loading-unloading events that emerge from the numerical calculations of the Lu–Klimas model. Firstly, after an initial transient, the sequence of burst-like unloading events seem to occur by clustering around an average field strength, i. e., on average, the increasing strength of the field reversal is more-or-less balanced by the annihilation of the magnetic field in the current sheet. Thus, we expect that the unloading events exhibit power law distributions in, for example, energy dissipation and burst duration. These results were essentially uncovered by Klimas et al. [66] for the special situation of very low driving rates and  $D_{\min} = 0$ .

We may also ask the question about the scale free invariant behavior for the complicated current sheet distri-

butions triggered by the instabilities associated with the anomalous resistivity for a single unloading event. Such small scale fluctuations again seem to self-consistently balance between the anomalous transport and field annihilation. Thus, we may again expect certain universal SOC behavior for these small scale fluctuations. To search for such behavior, Uritsky et al. [123] defined avalanches in terms of the size and duration of contiguous unstable grid points in the spatiotemporal domain. Power law statistics of the probability distributions of avalanche size and life time emerged from the statistics of their numerical calculations, Fig. 22. And, these exponents have been demonstrated to satisfy the standard exponent relations known for SOC processes [128]. There are other statistical verifications of SOC behavior from both the 1D and 2D dynamical modeling results, which we shall not discuss here.

More interestingly, Klimas et al. [69] recently demonstrated that the velocity fluctuations of their 2D simulation of the dynamical model exhibited both SOC and multifractal characteristics very similar to those expected for intermittent MHD turbulence; thus, corroborating with some of the basic concepts of interwoven connection and crossover phenomenon between SOC/FSOC and intermittent turbulence as espoused in the previous section.

### Future Directions

We have now come to perhaps the most important part of this article. As discussed in the introduction, the development of the modern concepts of dynamical complexity in space plasmas is very recent. Despite a flurry of activities, current investigations – both from the theoretical and observational points of view – have barely scratched the surface of what may be achieved and understood. For example, we need to know more about the complexity behavior of the plasmas from the kinetic point of view. We also need to know how the plasma particle distributions respond to the fluctuations of its turbulent environment. In addition, the global space environment is generally inhomogeneous with complex initial and boundary conditions. Thus we need to address processes that are statistically nonstationary and spatially inhomogeneous as well as those that evolve temporally and nonuniformly. We also need to understand how the complexity and multifractal phenomena of space plasmas influence the global dynamics of the space plasma environment.

Throughout the discourse, we have placed our emphasis of dynamical complexity in space plasmas on the basic underlying physics. We have come to realize that such fundamental ideas must come from the understanding of the complicated stochastic behavior arisen from the inter-



actions of the multitudes of coherent structures that are prevalent in the natural development of the dynamics of space plasmas. We learned that there were various types of invariant concepts related to coarse-graining, which characterize the primary phenomena that we are attempting to study, analyze and understand. We studied a suite of statistical methodology to analyze complexity, some of which are new and some are carryovers from other fields of complexity, including hydrodynamic turbulence.

We realized that most of the so-called theories that exist today in dynamical complexity of space plasmas are mainly phenomenological concepts and generally mimic those that were prevalent before the modern developments of complexity. We know that these theories cannot carry us too far as we know that they had not advanced the true understanding of complexity phenomena in the past. For this reason, we have broken the tradition in this article and tried not to continually compare our statistically acquired results and theoretically developed understandings with the prevalent classical ideas and numerology. Such comparisons abound in the literature and generally murk the basic concepts that we are trying to convey.

We have learned that most of the phenomena of complexity in space plasmas are scale invariant concepts and yet very little theory has been developed to try to explore these invariant characteristics. We are still stuck with the dogma that we must develop our theories from the “basic elemental equations”, while fully aware that the phenomena that we are trying to address generally have nothing to do with these equations directly. After all, nearly every invariant result gained from the coarse-graining concept breaks most of the symmetries and invariance properties of the basic elemental equations.

We have learned that the elemental dissipation such as classical resistivity and viscosity generally has nothing to do with the apparent dissipation that are everywhere in dynamical complexity. And yet every existing theoretical idea insists on including these almost irrelevant dissipative terms in its basic development.

We learned that intermittency generally involve large deviations from the mean and yet we insist on fitting every statistical result into neatly packed curves as if all physics and theories must result in smooth curves in parametric studies. It is time to bring in the ideas of Lebesgue measure, extreme value statistics and thermodynamics of rare events, as well as rank-ordered groupings and associated analyzes to our future developments of new theories and statistical methods

With enthusiasm, we look forward to the dawning of a new era from the future horizon when these new concepts and theoretical understandings finally establish their

footings in the modern development of dynamical complexity in space plasmas.

## Acknowledgments

The author is extremely grateful to Cheng-chin Wu for his unremitting collaboration and for many of the numerical results that are displayed in this article. He is indebted to Vassilis Angelopoulos, Roberto Bruno, Sandra Chapman, Abe Chian, Giuseppe Consolini, Marius Echim, Bogdan Hnat, Jack Jasperse, Charlie Kennel, Margy Kivelson, Alex Klimas, Hervé Lamy, Tony Lui, Eckart Marsch, Bill Matthaeus, John Podesta, Surja Sharma, Sunny Tam, Vadim Uritsky, Dimitri Vassiliadis, Zoltan Vörös, Nick Watkins, Jim Weygand, and Gary Zank for cooperative research and/or useful discussions. He is very appreciative to those who kindly allowed him to quote and reprint their published results and to Rich Chang for a critical reading of the manuscript and for his helpful suggestions. The author wishes to thank the Section Editor of Chaos and Complexity in Astrophysics for this encyclopedia volume, for inviting him to contribute this article and for his careful editing and useful suggestions. He is most grateful to members of his family for their constant encouragement and particularly to his wife, Monica, for her understanding and persistent spurring. Permissions from Elsevier, Springer Science and Business Media, the American Astrophysical Society, and the Italian Society of Physics and the liberal copyright policies of the American Physical Society, the American Institute of Physics, the American Geophysical Union, and the European Geophysical Union, which allowed the author to reprint figures and illustrations from their publications, are gratefully acknowledged. This project was partially supported by the AFOSR, NASA and NSF of the US government.

## Bibliography

### Primary Literature

1. Alexandrova O, Mangeney A, Maksimovic M, Cornilleau-Wehrin N, Bosqued JM, André M (2006) Alfvén vortex filaments observed in magnetosheath downstream of a quasi-perpendicular bow shock. *J Geophys Res* 111:A12208; doi:10.1029/2006JA011934
2. Angelopoulos V, Mukai T, Kokubun S (1999) Evidence for intermittency in Earth's plasma sheet and implications for self-organized criticality. *Phys Plasmas* 6:4161–4168
3. Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: an explanation of  $1/f$  noise. *Phys Rev Lett* 59:381–384
4. Baker DN, Klimas AJ, McPherson RL, Büchner J (1990) The evolution from weak to strong geomagnetic activity: An interpretation in terms of deterministic chaos. *Geophys Res Lett* 17:41–44
5. Balatoni J, Renyi A (1956) *Pub Math Inst Hung Acad Sci* 1:9

6. Benzi R, Ciliberto S, Tripiccone R, Baudet C, Massaioli F, Succi S (1993) Extended self-similarity in turbulent flows. *Phys Rev E* 48:R29–32
7. Biskamp D (1993) *Nonlinear magnetohydrodynamics*. Cambridge University Press, Cambridge
8. Biskamp D (2003) *Magnetohydrodynamic turbulence*. Cambridge University Press, Cambridge
9. Bruno R, Bavassano B, Pietropaolo E, Carbone V, Veltri P (1999) Effects of intermittency on interplanetary velocity and magnetic field fluctuations anisotropy. *Geophys Res Lett* 26:3185–3188
10. Bruno R, Carbone V, Veltri P, Pietropaolo E, Bavassano B (2001) Identifying intermittency events in the solar wind. *Planet Space Sci* 49:1201–1210
11. Bruno R, Carbone V, Chapman S, Hnat B, Noullez A, Sorriso-Valvo L (2007) Intermittent character of interplanetary magnetic field fluctuations. *Phys Plasmas* 14:032901
12. Burlaga LF (1991) Intermittent turbulence in the solar wind. *J Geophys Res* 96:5847–5851
13. Burlaga LF (1991) Multifractal structure of the interplanetary magnetic field: Voyager 2 observations near 25 AU, 1987–1988. *Geophys Res Lett* 18:69–72
14. Burlaga LF (1991) Multifractal structure of speed fluctuations in recurrent streams at 1 AU and near 6 AU. *Geophys Res Lett* 18:1651–1654
15. Burlaga LF (1993) Intermittent turbulence in large-scale velocity fluctuations at 1 AU near solar maximum. *J Geophys Res* 98:17467–17473
16. Carbone V (1993) Cascade model for intermittency in fully developed magneto-hydrodynamic turbulence. *Phys Rev Lett* 71:1546–1548
17. Carbone V (1994) Scaling exponents of the velocity structure functions in the interplanetary medium. *Ann Geophys* 12:585–590
18. Carbone V, Bruno R (1996) Cancellation exponents and multifractal scaling laws in the solar wind magnetohydrodynamic turbulence. *Ann Geophys* 14:777–785
19. Carbone V, Veltri P, Bruno R (1995) Experimental evidence for differences in the extended self-similarity scaling laws between fluid and magnetohydrodynamic turbulent flows. *Phys Rev Lett* 75:3110–3113
20. Carbone V, Bruno R, Veltri P (1996) Evidence for extended self-similarity in hydromagnetic turbulence. *Geophys Res Lett* 23:121–124
21. Castaing B, Gagne Y, Hopfinger EJ (1990) Velocity probability density functions of high Reynolds number turbulence. *Physica D* 46:177–200
22. Chang T (1992) Low-dimensional behavior and symmetry breaking of stochastic systems near criticality – Can these effects be observed in space and in the laboratory? *IEEE Trans Plasma Sci* 20:691–694
23. Chang T (1998) Sporadic localized reconnections and multi-scale intermittent turbulence in the magnetotail. In: Horwitz JL, Gallagher DI, Peterson WK (eds) *Geospace mass and energy flow*. Geophysical Monograph, vol 104. AGU, Washington, pp 193–199
24. Chang T (1999) Self-organized criticality, multifractal spectra, sporadic localized reconnections and intermittent turbulence in the magnetotail. *Phys Plasmas* 6:4137–4145
25. Chang T (2001) An example of resonances, coherent structures and topological phase transitions – the origin of the low frequency broadband spectrum in the auroral zone. *Nonlinear Process Geophys* 8:175–179
26. Chang T (2001) Colloid-like behavior and topological phase transitions in space plasmas: intermittent low frequency turbulence in the auroral zone. *Phys Scr* T89:80–83
27. Chang T, Wu CC (2007) Dynamical complexity, intermittent turbulence, coarse-grained dissipation, criticality and multifractal processes. In: Shaikh D (ed) *Turbulence and nonlinear processes in astrophysical plasmas*. AIP Proc 932:161–166
28. Chang T, Wu CC (2008) Rank-ordered multifractal spectrum for intermittent fluctuations. *Phys Rev E* 77:045401(R)
29. Chang T, Hankey A, Stanley HE (1973) Generalized scaling hypothesis in multicomponent systems, vol 1. Classification of critical points by order and scaling at tricritical points. *Phys Rev B* 8:346–364
30. Chang T, Nicoll JF, Young JF (1978) A closed-form differential renormalization-group generator for critical dynamics. *Phys Lett A* 67:287–290
31. Chang T, Vvedensky DD, Nicoll JF (1992) Differential renormalization-group generators for static and dynamic critical phenomena. *Phys Reports* 217:279–362
32. Chang T, Tam SWY, Wu CC, Consolini G (2003) Complexity, forced and/or self-organized criticality, and topological phase transitions in space plasmas. *Space Sci Rev* 107:425–445
33. Chang T, Tam SWY, Wu CC (2004) Complexity induced anisotropic bimodal intermittent turbulence in space plasmas. *Phys Plasmas* 11:1287–1299
34. Chang T, Tam SWY, Wu CC (2006) Complexity in space plasmas – a brief review. *Space Sci Rev* 122:281–291
35. Chapman SC, Watkins NW, Dendy RO, Helander P, Rowlands G (1998) A simple avalanche model as an analogue for the magnetospheric activity. *Geophys Res Lett* 25:2397–2400
36. Chapman SC, Dendy RO, Rowlands G (1999) A sandpile model with dual scaling regimes for laboratory, space, and astrophysical plasmas. *Phys Plasmas* 6:4169–4177
37. Consolini G (1997) Sandpile cellular automata and the magnetospheric dynamics. In: Aiello S, Iucci N, Sironi G, Treves A, Villante U (ed) *Cosmic Physics in the Year 2000*. Proc VIII GIFCO Conference, Società Italiana di Fisica, Bologna, pp 123–126
38. Consolini G, Chang T (2001) Magnetic field topology and criticality in geotail dynamics: Relevance to substorm phenomena. *Space Sci Rev* 95:309
39. Consolini G, Chang T (2002) Complexity, magnetic field topology, criticality, and metastability in magnetotail dynamics. *J Atmos Solar Terr Phys* 64:541
40. Consolini G, Lui ATY (2000) Symmetry Breaking and Nonlinear Wave-Wave Interaction in Current Disruption: Possible evidence for a phase Transition. In: Ohtani S et al (eds) *Magnetospheric Current Systems*. Geophysical Monograph, vol 118. AGU, Washington, p 395
41. Consolini G, Marcucci MF, Candidi M (1996) Multifractal structure of auroral electrojet index data. *Phys Rev Lett* 76:4082–4085
42. Consolini G, Chang T, Lui ATY (2005) Complexity and topological disorder in the Earth's magnetotail dynamics. In: Sharma AS, Kaw PK (eds) *Nonequilibrium phenomena in plasmas*. Springer, Dordrecht, pp 51–69
43. Daubechies I (1992) *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, Philadelphia, p 357

44. Echim MM, Lamy H, Chang T (2007) Multipoint observations of intermittency in the cusp regions. *Nonlinear Processes in Geophysics* 14:525–534
45. Einaudi G, Velli M (1999) The distribution of flares, statistics of magnetohydrodynamic turbulence and coronal heating. *Phys Plasmas* 6:4146–4153
46. Farge M (1992) Wavelet transforms and their applications to turbulence. *Ann Rev Fluid Mech* 24:395–457
47. Feynman J, Ruzmaikin A (1994) Distributions of the interplanetary field revisited. *J Geophys Res* 99:17645–17651
48. Forman M, Burlaga LF (2003) Exploring the Castaing distribution function to study intermittence in the solar wind at L1 in June 2000. In: Velli M, Bruno R, Malara F (eds) *Solar wind ten*. AIP Conf Proc 679:554–557
49. Forster D, Nelson DR, Stephen MJ (1977) Large distance and long time properties of a randomly stirred fluid. *Phys Rev A* 16:732–749
50. Grassberger P, Procaccia I (1983) Characterization of strange attractors. *Phys Rev Lett* 50:346–349
51. Grossmann S, Lohse D, Reeh A (1997) Application of extended self-similarity in turbulence. *Phys Rev E* 56:5473–5478
52. Haar A (1910) Zur Theorie der Orthogonalen Functionensysteme. *Math Ann* 69:331–371
53. Halperin BI, Hohenberg PC, Ma SK (1972) Calculation of dynamical critical properties using Wilson's expansion methods. *Phys Rev Lett* 29:1548–1551
54. Halsey T, Jensen MH, Kadanoff LP, Procaccia I, Schraiman BI (1986) Fractal measures and their singularities: The characterization of strange sets. *Phys Rev A* 33:1141–1151
55. Hentschel HGE, Procaccia I (1983) The infinite number of generalized dimensions of fractals and strange attractors. *Physica* 8D:435–444
56. Hnat B, Chapman SC, Rowlands G, Watkins NW, Farrell WM (2002) Finite size scaling in the solar wind magnetic field energy density as seen by WIND. *Geophys Res Lett* 29:1446; doi:10.1029/2001GL014587
57. Hnat B, Chapman SC, Rowlands G, Watkins NW, Freeman MP (2003) Scaling in long term data sets of geomagnetic indices and solar wind  $\varepsilon$  as seen by WIND spacecraft. *Geophys Res Lett* 30:2174; doi:10.1029/2003GL018209
58. Horbury TA, Balogh A, Forsyth RJ, Smith EJ (1995) Ulysses magnetic field observations of fluctuations within polar coronal flows. *Ann Geophys* 13:105–107
59. Horbury TA, Balogh A, Forsyth RJ, Smith EJ (1995) Observations of evolving turbulence in the polar solar wind. *Geophys Res Lett* 22:3401–3404
60. Horbury TA, Balogh A, Forsyth RJ, Smith EJ (1997) Ulysses observations of intermittent heliospheric turbulence. *Adv Space Phys* 19:847–850
61. Hosino M, Nishida A, Yamamoto T, Kokubun S (1994) Turbulent magnetic field in the distant magnetotail: Bottom-up process of plasmoid formation? *Geophys Res Lett* 21:2935–2938
62. Hwa T, Kardar M (1992) Avalanches, hydrodynamics, and discharge events in models of sandpiles. *Phys Rev A* 45:7002–7023
63. Jensen HJ (1998) Self-organized criticality – Emergent complex behavior in physical and biological systems. Cambridge University Press, Cambridge
64. Jiang GS, Wu CC (1999) A high-order WENO finite difference scheme for the equations of ideal magnetohydrodynamics. *J Comp Phys* 150:561–594
65. Klimas AJ, Baker DN, Roberts DA, Fairfield DH, Büchner J (1992) A nonlinear dynamical analogue model of geomagnetic activity. *J Geophys Res* 97:12253–12266
66. Klimas AJ, Valdivia JA, Vassiliadis D, Baker DN, Hesse M, Takalo J (2000) Self-organized criticality in the substorm phenomenon and its relation to localized reconnection in the magnetospheric plasma sheet. *J Geophys Res* 105:18765–18780
67. Klimas AJ, Uritsky VM, Vassiliadis D, Baker DN (2004) Reconnection and scale-free avalanching in a driven current-sheet model. *J Geophys Res* 109:A02218; doi:10.1029/2003JA010036
68. Klimas AJ, Uritsky VM, Vassiliadis D, Baker DN (2005) A mechanism for the loading-unloading substorm cycle missing in MHD global magnetospheric simulation models. *Geophys Res Lett* 32:L14108; doi:10.1029/2005GL022916
69. Klimas AJ, Uritsky VM, Paczuski M (2007) Self-organized criticality and intermittent turbulence in an MHD current sheet with a threshold instability. arXiv:astro-ph/0701486v2
70. Langmuir I (1928) Oscillations in ionized gases. *Proc Natl Acad Sci* 14:628
71. Leubner LP (2004) Fundamental issues on kappa-distributions in space plasmas and interplanetary proton distributions. *Phys Plasmas* 11:1308–1316
72. Leubner MP, Vörös Z (2005) A nonextensive entropy approach to solar wind intermittency. *Astrophys J* 618:547–555
73. Lu ET (1995) Avalanches in continuum driven dissipative systems. *Phys Rev Lett* 74:2511–2514
74. Lu ET, Hamilton RJ (1991) Avalanches and the distribution of solar flares. *Astrophys J* 380:L89–L92
75. Lui ATY (1996) Current disruption in the Earth's magnetosphere: observations and models. *J Geophys Res* 101:13067–13088
76. Lui ATY (2001) Multifractal and intermittent nature of substorm-associated magnetic turbulence in the magnetotail. *J Atmos Solar Terr Phys* 63:1379–1385
77. Lui ATY (2002) Multiscale phenomena in the near-Earth magnetosphere. *J Atmos Solar Terr Phys* 64:125
78. Lui ATY, Lopez RE, Krimigis SM, McEntire RW, Zanetti LJ, Potemra TA (1988) A case study of magnetotail current sheet disruption and diversion. *Geophys Res Lett* 15:721–724
79. Lui ATY, Chang CL, Mankofsky A, Wong HK, Winske D (1991) A cross-field current instability for substorm expansions. *J Geophys Res* 96:11389–11401
80. Lui ATY, Liou K, Newell PT, Meng CI, Ohtani SI, Kokubun S, Ogino T, Brittner M, Parks G (1998) Plasma and magnetic flux transport associated with auroral breakups. *Geophys Res Lett* 25:4059–4062
81. Lui ATY, Chapman SC, Liou K, Newell PT, Meng CI, Brittner M, Parks GK (2000) Is the dynamic magnetosphere an avalanching system? *Geophys Res Lett* 27:911–914
82. Lui ATY, Zheng Y, Zhang Y, Rème H, Dunlop MW, Gustafsson G, Mende SB, Mouikis C, Kistler LM (2006) Cluster observation of plasma flow reversal in the magnetotail during a substorm. *Ann Geophys* 24:2005–2013
83. Mandelbrot B (1977) *Fractals: Form, chance and dimension*. Freeman, San Francisco

84. Marsch E, Liu S (1993) Structure functions and intermittency of velocity fluctuations in the inner solar wind. *Ann Geophys* 11:227–238
85. Marsch E, Tu CY (1994) Non-Gaussian probability distributions of solar wind fluctuations. *Ann Geophys* 12:1127–1138
86. Marsch E, Tu CY (1997) Intermittency, non-Gaussian statistics and fractal scaling of MHD fluctuations in the solar wind. *Nonlinear Process Geophys* 4:101–124
87. Marsch E, Tu CY, Rosenbauer H (1996) Multifractal scaling of the kinetic energy flux in solar wind turbulence. *Ann Geophys* 14:259–269
88. Matthaeus WH, Goldstein ML (1982) Measurements of the rugged invariants of magnetohydrodynamic turbulence in the solar wind. *J Geophys Res* 87:6011–6028
89. Matthaeus WH, Goldstein ML, Roberts DA (1990) Evidence for the presence of quasi-two-dimensional nearly incompressible fluctuations in the solar wind. *J Geophys Res* 95:20673–20683
90. Matthaeus WH, Qin G, Bieber JW, Zank GP (2003) Nonlinear collisionless perpendicular diffusion of charged particles. *Astrophys J Lett* 590:L53–L56
91. Meneveau C (1991) Analysis of turbulence in the orthonormal wavelet representation. *J Fluid Mech* 232:469–521
92. Meneveau C, Sreenivasan KR (1987) Simple multifractal cascade model for fully developed turbulence. *Phys Rev Lett* 59:1424–1427
93. Milovanov AV, Zelenyi LM, Zimbardo G (1996) Fractal structures and power law spectra in the distant Earth's magnetotail. *J Geophys Res* 101:19903–19910
94. Morlet J (1983) Sampling theory and wave propagation. In: Chen CH (ed) *Issues on acoustic signal/image processing and recognition*, NATO ASI, vol F1. Springer, Berlin
95. Müller WC, Biskamp D (2002) Scaling properties of three-dimensional magnetohydrodynamic turbulence. *Phys Rev Lett* 84:475–478
96. Ott E (1993) *Chaos in dynamical systems*. Cambridge University Press, New York
97. Pagel C, Balogh A (2001) A study of magnetic fluctuations and their anomalous scaling in the solar wind: the Ulysses fast-latitude scan. *Nonlinear Process Geophys* 8:313–330
98. Parisi G, Frisch U (1985) On the singularity structure of fully developed turbulence. In: Ghil M, Benzi R, Parisi G (eds) *Turbulence and predictability in geophysical fluid dynamics*. Proc Intern School of Physics, 'Erico Fermi' 1983, Varenna, Italy. North Holland, Amsterdam, pp 84–87
99. Pinheiro MJ (2007) Plasmas: the genesis of the word. *arXiv:physics/0703260v1*
100. Politano H, Pouquet A, Carbone V (1998) Determination of anomalous exponents of structure functions in two-dimensional magnetohydrodynamic turbulence. *Europhys Lett* 43:516–521
101. Roberts DA, Baker DN, Klimas AJ, Bargatze LF (1991) Indications of low dimensionality in magnetospheric dynamics. *Geophys Res Lett* 18:151–154
102. Rusmaikin A, Lyannaya IR, Styashkin VA, Yeroshenko E (1993) The spectrum of the interplanetary magnetic field near 1.3 AU. *J Geophys Res* 98:13303–13306
103. Shan LH, Hansen P, Goertz CK, Smith RA (1991) Chaotic appearance of the AE index. *Geophys Res Lett* 18:147–150
104. Shan LH, Goertz CK, Smith RA (1991) On the embedding-dimension analysis of AE and AL time series. *Geophys Res Lett* 18:1647–1650
105. Sharma AS, Vassiliadis DV, Papadopoulos K (1993) Reconstruction of low-dimensional magnetospheric dynamics by singular spectrum analysis. *Geophys Res Lett* 20:335–338
106. Sitnov GL, Sharma AS, Papadopoulos K, Vassiliadis D, Valdivia JA, Klimas AJ, Baker DN (2000) Phase transition-like behavior of the magnetosphere during substorms. *J Geophys Res* 105:12955–12974
107. Sorriso-Valvo L, Carbone V, Veltri P, Consolini G, Bruno R (1999) Intermittency in the solar wind turbulence through probability distribution functions of fluctuations. *Geophys Res Lett* 26:1801–1804
108. Sorriso-Valvo L, Carbone V, Giuliani P, Veltri P, Bruno R, Antonini V, Martines E (2001) Intermittency in plasma turbulence. *Planet Space Sci* 49:1193–1200
109. Sundkvist D, Krasnoselskikh V, Shukla PK, Vaivads A, André M, Buchert S, Rème H (2005) In situ multi-satellite detection of coherent vortices as a manifestation of Alfvénic turbulence. *Nature* 436:825–828
110. Takalo J, Timonen J, Klimas AJ, Valdivia JA, Vassiliadis D (2001) A coupled map as a model of the dynamics of the magnetotail current sheet. *J Atmos Solar Terr Phys* 63:1407–1414
111. Tam SWY, Chang T (2005) Energization of ions by bimodal intermittent fluctuations. In: Lui ATY, Kamide Y, Consolini G (eds) *Multiscale coupling of sun-earth processes*. Elsevier, Netherlands, pp 375
112. Tam SWY, Chang T, Kintner PM, Klatt E (2005) Intermittency analyses on the SIERRA measurements of the electric field fluctuations in the auroral zone. *Geophys Res Lett* 32:L05109; doi:10.1029/2004GL021445
113. Taylor GI (1938) The spectrum of turbulence. *Proc R Soc Lond A* 164:476–490
114. Taylor JB (1974) Relaxation of toroidal plasma and generation of reverse magnetic fields. *Phys Rev Lett* 33:1139–1141
115. Tetreault D (1991) Theory of electric fields in the auroral acceleration region. *J Geophys Res* 96:3549–3563
116. Tetreault D (1992) Turbulent relaxation of magnetic fields, 1. Coarse-grained dissipation and reconnection. *J Geophys Res* 97:8531–8540
117. Tetreault D (1992) Turbulent relaxation of magnetic fields, 2. Self-organization and intermittency. *J Geophys Res* 97:8541–8547
118. Tonks L (1967) *Am J Phys* 35:857–858
119. Tonks L, Langmuir I (1929) Oscillations in ionized gases. *Phys Rev* 33:195–210
120. Tu CY, Marsch E (1995) MHD structures, waves and turbulence in the solar wind – observations and theories. Kluwer, Dordrecht
121. Tu CY, Marsch E, Rosenbauer H (1996) An extended structure function model and its application to the analysis of solar wind intermittency properties. *Ann Geophys* 14:270–285
122. Uritsky VM, Klimas AJ (2005) Hysteresis-controlled instability waves in a scale-free driven current sheet model. *Nonlinear Process Geophys* 12:827–833
123. Uritsky VM, Klimas AJ, Valdivia JA, Vassiliadis D, Baker DN (2001) Stable critical behavior and fast field annihilation in a magnetic field reversal model. *J Atmos Solar Terr Phys* 63:1425–1433



124. Uritsky VM, Klimas AJ, Vassiliadis D, Chua D, Parks G (2002) Scale-free statistics of spatiotemporal auroral emissions as depicted by POLAR UVI images: Dynamic magnetosphere is an avalanching system. *J Geophys Res* 107:1426; doi:10.1029/2001JA000281
125. Uritsky VM, Klimas AJ, Vassiliadis D (2002) Multiscale dynamics and robust critical scaling in a continuum current sheet model. *Phys Rev E* 65:046113
126. Uritsky VM, Paczuski M, Davila JM, Jones SI (2007) Coexistence of self-organized criticality and intermittent turbulence in the solar corona. *Phys Rev Lett* 99:025001
127. Vahnstein S, Sreenivasan K, Pierrehumbert R, Kashyap V, Juneja A (1994) Scaling exponents for turbulence and other random processes and their relationships with multifractal structure. *Phys Rev E* 50:1823–1835
128. Vespignani A, Zapperi S (1998) How self-organized criticality works: A unified mean-field picture. *Phys Rev E* 57:6345–6362
129. Vörös Z (2000) On multifractality of high-latitude geomagnetic fluctuations. *Ann Geophys* 18:1273–1283
130. Vörös Z, Baumjohann W, Nakamura R, Runov A, Zhang TL, Volwerk M, Eichelberger HU, Balogh A, Horbury TS, Glassmeier KH, Klecker B, Rème H (2003) Multi-scale magnetic field intermittence in the plasma sheet. *Ann Geophys* 21:1955–1964
131. Vörös Z, Baumjohann W, Nakamura R, Runov A, Volwerk M, Zhang TL, Balogh A (2004) Wavelet analysis of magnetic turbulence in the Earth's plasma sheet. *Phys Plasmas* 11:1333–1338
132. Vörös Z, Baumjohann W, Nakamura R, Runov A, Volwerk M, Schwarzl H, Balogh A, Rème H (2005) Dissipation scales in the Earth's plasma sheet estimated from cluster measurements. *Nonlinear Process Geophys* 12:725–732
133. Watkins NW, Chapman SC, Dendy RO, Rowlands G (1999) Robustness of collective behaviour in a strongly driven avalanche model: magnetospheric implications. *Geophys Res Lett* 26:2617–2620
134. Weygand JM, Kivelson MG, Khurana KK, Schwarzl HK, Thompson SM, McPherron RL, Balogh A, Kistler LM, Goldstein ML, Borovsky J, Roberts DA (2005) Plasma sheet turbulence observed by Cluster, vol II. *J Geophys Res* 110:A01205; doi:10.1029/2004JA010581
135. Wilson KG, Kogut J (1994) The renormalization group and the  $\varepsilon$  expansion. *Phys Reports* 12C:75–200
136. Wu CC, Chang T (2000) 2D MHD simulation of the emergence and merging of coherent structures. *Geophys Res Lett* 27:863–866
137. Wu CC, Chang T (2001) Further study of the dynamics of two-dimensional MHD coherent structures – a large scale simulation. *J Atmos Solar Terr Phys* 63:1447–1453
138. Wu CC, Chang T (2005) Intermittent turbulence in 2D MHD simulation. In: Lui ATY, Kamide Y, Consolini G (eds) Multiscale coupling of sun-earth processes. Elsevier, Netherlands, p 321
139. Yordanova E, Grezasiak M, Wernik AW, Popielawska B, Stasiewicz K (2005) Multifractal structure of turbulence in the magnetospheric cusp. *Ann Geophys* 22:2431–2440
- Chang T, Chapman S, Klimas A (eds) (2001) Forced and/or self-organized criticality (FSOC) in space plasmas. *J Atmos Solar Terr* 63:1359–1453
- Forster D (1975) Hydrodynamic fluctuations, broken symmetry, and correlation functions. Benjamin, Reading
- Frisch U (1995) Turbulence. Cambridge University Press, Cambridge
- Lavenda BH (1995) Thermodynamics of Extremes. Albion Publishers, Chichester
- Leubner MP, Baumjohann W, Chian ACL (eds) (2006) Advances in space environment research, vol 2. *Space Sci Rev* 122:1–337
- Lui ATY, Kamide Y, Consolini G (eds) (2005) Multiscale coupling of sun-earth processes. Elsevier, the Netherlands
- Ma SK (1976) Modern theory of critical phenomena. Benjamin, Reading
- Sharma AS, Kaw PK (eds) (2005) Nonequilibrium phenomena in plasmas. Astrophysics and Space Science Library, vol 321. Springer, Dordrecht, pp 1–340
- Sornette D (2000) Critical phenomena in natural sciences; Chaos, fractals, selforganization and disorder: concepts and tools. Springer, Berlin

## Spectral Theory of Dynamical Systems

MARIUSZ LEMAŃCZYK

Faculty of Mathematics and Computer Science,  
Nicolaus Copernicus University, Toruń, Poland

### Article Outline

Glossary

Definition of the Subject

Introduction

Maximal Spectral Type of a Koopman Representation,  
Alexeyev's Theorem

Spectral Theory of Weighted Operators

The Multiplicity Function

Rokhlin Cocycles

Rank-1 and Related Systems

Spectral Theory of Dynamical Systems of Probabilistic  
Origin

Inducing and Spectral Theory

Special Flows and Flows on Surfaces, Interval Exchange  
Transformations

Future Directions

Acknowledgments

Bibliography

### Books and Reviews

- Bruno R, Carbone V (2005) The solar wind as a turbulence laboratory. *Living Rev Solar Phys* 2:4; <http://www.livingreviews.org/lrsp-2005-4>. Accessed 15 June 2007

### Glossary

**Spectral decomposition of a unitary representation** If  $\mathcal{U} = (U_a)_{a \in \mathbb{A}}$  is a continuous unitary representation of a locally compact second countable (l.c.s.c.)



Abelian group  $\mathbb{A}$  in a separable Hilbert space  $H$  then a decomposition  $H = \bigoplus_{i=1}^{\infty} \mathbb{A}(x_i)$  is called *spectral* if  $\sigma_{x_1} \gg \sigma_{x_2} \gg \dots$  (such a sequence of measures is also called *spectral*); here  $\mathbb{A}(x) := \overline{\text{span}}\{U_a x : a \in \mathbb{A}\}$  is called the *cyclic space* generated by  $x \in H$  and  $\sigma_x$  stands for the spectral measure of  $x$ .

### Maximal spectral type and the multiplicity function of $\mathcal{U}$

The *maximal spectral type*  $\sigma_{\mathcal{U}}$  of  $\mathcal{U}$  is the type of  $\sigma_{x_1}$  in any spectral decomposition of  $H$ ; the *multiplicity function*  $M_{\mathcal{U}}: \widehat{\mathbb{A}} \rightarrow \{1, 2, \dots\} \cup \{+\infty\}$  is defined  $\sigma_{\mathcal{U}}$ -a.e. and  $M_{\mathcal{U}}(\chi) = \sum_{i=1}^{\infty} 1_{Y_i}(\chi)$ , where  $Y_1 = \widehat{\mathbb{A}}$  and  $Y_i = \text{supp } d\sigma_{x_i}/d\sigma_{x_1}$  for  $i \geq 2$ .

A representation  $\mathcal{U}$  is said to have *simple spectrum* if  $H$  is reduced to a single cyclic space. The multiplicity is *uniform* if there is only one essential value of  $M_{\mathcal{U}}$ . The essential supremum of  $M_{\mathcal{U}}$  is called the *maximal spectral multiplicity*.  $\mathcal{U}$  is said to have *discrete spectrum* if  $H$  has an orthonormal base consisting of eigenvectors of  $\mathcal{U}$ ;  $\mathcal{U}$  has *singular* (Haar, absolutely continuous) spectrum if the maximal spectral type of  $\mathcal{U}$  is singular with respect to (equivalent to, absolutely continuous with respect to) a Haar measure of  $\widehat{\mathbb{A}}$ .

**Koopman representation of a dynamical system  $\mathcal{T}$**  Let  $\mathbb{A}$  be a l.c.s.c. (and not compact) Abelian group and  $\mathcal{T}: a \mapsto T_a$  a representation of  $\mathbb{A}$  in the group  $\text{Aut}(X, \mathcal{B}, \mu)$  of (measure-preserving) automorphisms of a standard probability Borel space  $(X, \mathcal{B}, \mu)$ . The *Koopman representation*  $\mathcal{U} = \mathcal{U}_{\mathcal{T}}$  of  $\mathcal{T}$  in  $L^2(X, \mathcal{B}, \mu)$  is defined as the unitary representation  $a \mapsto U_{T_a} \in U(L^2(X, \mathcal{B}, \mu))$ , where  $U_{T_a}(f) = f \circ T_a$ .

**Ergodicity, weak mixing, mild mixing, mixing and rigidity of  $\mathcal{T}$**  A measure-preserving  $\mathbb{A}$ -action  $\mathcal{T} = (T_a)_{a \in \mathbb{A}}$  is called *ergodic* if  $\chi_0 \equiv 1 \in \widehat{\mathbb{A}}$  is a simple eigenvalue of  $\mathcal{U}_{\mathcal{T}}$ . It is *weakly mixing* if  $\mathcal{U}_{\mathcal{T}}$  has a continuous spectrum on the subspace  $L_0^2(X, \mathcal{B}, \mu)$  of zero mean functions.  $\mathcal{T}$  is said to be *rigid* if there is a sequence  $(a_n)$  going to infinity in  $\mathbb{A}$  such that the sequence  $(U_{T_{a_n}})$  goes to the identity in the strong (or weak) operator topology;  $\mathcal{T}$  is said to be *mildly mixing* if it has no non-trivial rigid factors. We say that  $\mathcal{T}$  is *mixing* if the operator equal to zero is the only limit point of  $\{U_{T_a}|_{L_0^2(X, \mathcal{B}, \mu)} : a \in \mathbb{A}\}$  in the weak operator topology.

**Spectral disjointness** Two  $\mathbb{A}$ -actions  $S$  and  $\mathcal{T}$  are called *spectrally disjoint* if the maximal spectral types of their Koopman representations  $\mathcal{U}_{\mathcal{T}}$  and  $\mathcal{U}_S$  on the corresponding  $L_0^2$ -spaces are mutually singular.

**SCS property** We say that a Borel measure  $\sigma$  on  $\widehat{\mathbb{A}}$  satisfies the *strong convolution singularity* property (SCS property) if, for each  $n \geq 1$ , in the disintegration (given by the map  $(\chi_1, \dots, \chi_n) \mapsto \chi_1 \cdot \dots \cdot \chi_n$ )

$\sigma^{\otimes n} = \int_{\widehat{\mathbb{A}}} \nu_{\chi} d\sigma^{(n)}(\chi)$  the conditional measures  $\nu_{\chi}$  are atomic with exactly  $n!$  atoms ( $\sigma^{(n)}$  stands for the  $n$ th convolution  $\sigma * \dots * \sigma$ ). An  $\mathbb{A}$ -action  $\mathcal{T}$  satisfies the SCS property if the maximal spectral type of  $\mathcal{U}_{\mathcal{T}}$  on  $L_0^2$  is a type of an SCS measure.

**Kolmogorov group property** An  $\mathbb{A}$ -action  $\mathcal{T}$  satisfies the *Kolmogorov group property* if  $\sigma_{\mathcal{U}_{\mathcal{T}}} * \sigma_{\mathcal{U}_{\mathcal{T}}} \ll \sigma_{\mathcal{U}_{\mathcal{T}}}$ .

**Weighted operator** Let  $T$  be an ergodic automorphism of  $(X, \mathcal{B}, \mu)$  and  $\xi: X \rightarrow \mathbb{T}$  be a measurable function. The (unitary) operator  $V = V_{\xi, T}$  acting on  $L^2(X, \mathcal{B}, \mu)$  by the formula  $V_{\xi, T}(f)(x) = \xi(x)f(Tx)$  is called a *weighted operator*.

**Induced automorphism** Assume that  $T$  is an automorphism of a standard probability Borel space  $(X, \mathcal{B}, \mu)$ . Let  $A \in \mathcal{B}$ ,  $\mu(A) > 0$ . The *induced automorphism*  $T_A$  is defined on the conditional space  $(A, \mathcal{B}_A, \mu_A)$ , where  $\mathcal{B}_A$  is the trace of  $\mathcal{B}$  on  $A$ ,  $\mu_A(B) = \mu(B)/\mu(A)$  for  $B \in \mathcal{B}_A$  and  $T_A(x) = T^{k_A(x)}x$ , where  $k_A(x)$  is the smallest  $k \geq 1$  for which  $T^k x \in A$ .

**AT property of an automorphism** An automorphism  $T$  of a standard probability Borel space  $(X, \mathcal{B}, \mu)$  is called *approximatively transitive* (AT for short) if for every  $\varepsilon > 0$  and every finite set  $f_1, \dots, f_n$  of non-negative  $L^1$ -functions on  $(X, \mathcal{B}, \mu)$  we can find  $f \in L^1(X, \mathcal{B}, \mu)$  also non-negative such that  $\|f_i - \sum_j \alpha_{ij} f \circ T^{n_j}\|_{L^1} < \varepsilon$  for all  $i = 1, \dots, n$  (for some  $\alpha_{ij} \geq 0, n_j \in \mathbb{N}$ ).

**Cocycles and group extensions** If  $T$  is an ergodic automorphism,  $G$  is a l.c.s.c. Abelian group and  $\varphi: X \rightarrow G$  is measurable then the pair  $(T, \varphi)$  generates a *cocycle*  $\varphi^{(\cdot)}(\cdot): \mathbb{Z} \times X \rightarrow G$ , where

$$\varphi^{(n)}(x) = \begin{cases} \varphi(x) + \dots + \varphi(T^{n-1}x) & \text{for } n > 0, \\ 0 & \text{for } n = 0, \\ -(\varphi(T^n x) + \dots + \varphi(T^{-1}x)) & \text{for } n < 0. \end{cases}$$

(That is  $(\varphi^{(n)})$  is a standard 1-cocycle in the algebraic sense for the  $\mathbb{Z}$ -action  $n(f) = f \circ T^n$  on the group of measurable functions on  $X$  with values in  $G$ ; hence the function  $\varphi: X \rightarrow G$  itself is often called a *cocycle*.)

Assume additionally that  $G$  is compact. Using the cocycle  $\varphi$  we define a *group extension*  $T_{\varphi}$  on  $(X \times G, \mathcal{B} \otimes \mathcal{B}(G), \mu \otimes \lambda_G)$  ( $\lambda_G$  stands for Haar measure of  $G$ ), where  $T_{\varphi}(x, g) = (Tx, \varphi(x) + g)$ .

**Special flow** Given an ergodic automorphism  $T$  on a standard probability Borel space  $(X, \mathcal{B}, \mu)$  and a positive integrable function  $f: X \rightarrow \mathbb{R}^+$  we put

$$X^f = \{(x, t) \in X \times \mathbb{R} : 0 \leq t < f(x)\}, \\ \mathcal{B}^f = \mathcal{B} \otimes \mathcal{B}(\mathbb{R})|_{X^f},$$

and define  $\mu^f$  as normalized  $\mu \otimes \lambda_{\mathbb{R}}|_{X^f}$ . By a *special flow* we mean the  $\mathbb{R}$ -action  $T^f = (T_t^f)_{t \in \mathbb{R}}$  under which a point  $(x, s) \in X^f$  moves vertically with the unit speed, and once it reaches the graph of  $f$ , it is identified with  $(Tx, 0)$ .

**Markov operator** A linear operator  $J: L^2(X, \mathcal{B}, \mu) \rightarrow L^2(Y, \mathcal{C}, \nu)$  is called *Markov* if it sends non-negative functions to non-negative functions and  $J1 = J^*1 = 1$ .

**Unitary actions on Fock spaces** If  $H$  is a separable Hilbert space then by  $H^{\odot n}$  we denote the subspace of  $n$ -tensors of  $H^{\otimes n}$  symmetric under all permutations of coordinates,  $n \geq 1$ ; then the Hilbert space  $F(H) := \bigoplus_{n=0}^{\infty} H^{\odot n}$  is called a *symmetric Fock space*. If  $V \in U(H)$  then  $F(V) := \bigoplus_{n=0}^{\infty} V^{\odot n} \in U(F(H))$  where  $V^{\odot n} = V^{\otimes n}|_{H^{\odot n}}$ .

## Definition of the Subject

Spectral theory of dynamical systems is a study of special unitary representations, called Koopman representations (see the glossary). Invariants of such representations are called spectral invariants of measure-preserving systems. Together with the entropy, they constitute the most important invariants used in the study of measure-theoretic intrinsic properties and classification problems of dynamical systems as well as in applications. Spectral theory was originated by von Neumann, Halmos and Koopman in the 1930s. In this article we will focus on recent progresses in the spectral theory of finite measure-preserving dynamical systems.

## Introduction

Throughout  $\mathbb{A}$  denotes a non-compact l.c.s.c. Abelian group ( $\mathbb{A}$  will be most often  $\mathbb{Z}$  or  $\mathbb{R}$ ). The assumption of second countability implies that  $\mathbb{A}$  is metrizable,  $\sigma$ -compact and the space  $C_0(\mathbb{A})$  is separable. Moreover the dual group  $\widehat{\mathbb{A}}$  is also l.c.s.c. Abelian.

## General Unitary Representations

We are interested in *unitary*, that is with values in the unitary group  $U(H)$  of a Hilbert space  $H$ , (weakly) continuous representations  $V: \mathbb{A} \ni a \mapsto V_a \in U(H)$  of such groups (the scalar valued maps  $a \mapsto \langle V_a x, y \rangle$  are continuous for each  $x, y \in H$ ).

Let  $H = L^2(\widehat{\mathbb{A}}, \mathcal{B}(\widehat{\mathbb{A}}), \mu)$ , where  $\mathcal{B}(\widehat{\mathbb{A}})$  stands for the  $\sigma$ -algebra of Borel sets of  $\widehat{\mathbb{A}}$  and  $\mu \in M^+(\widehat{\mathbb{A}})$  (whenever  $X$  is a l.c.s.c. space, by  $M(X)$  we denote the set of complex Borel measures on  $X$ , while  $M^+(X)$  stands for the

subset of positive (finite) measures). Given  $a \in \mathbb{A}$ , for  $f \in L^2(\widehat{\mathbb{A}}, \mathcal{B}(\widehat{\mathbb{A}}), \mu)$  put

$$V_a^\mu(f)(\chi) = i(a)(\chi) \cdot f(\chi) = \chi(a) \cdot f(\chi) \quad (\chi \in \widehat{\mathbb{A}}),$$

where  $i: \mathbb{A} \rightarrow \widehat{\widehat{\mathbb{A}}}$  is the canonical Pontriagin isomorphism of  $\mathbb{A}$  with its second dual. Then  $V^\mu = (V_a^\mu)_{a \in \mathbb{A}}$  is a unitary representation of  $\mathbb{A}$ . Since  $C_0(\widehat{\mathbb{A}})$  is dense in  $L^2(\widehat{\mathbb{A}}, \mu)$ , the latter space is separable. Therefore also direct sums  $\bigoplus_{i=1}^{\infty} V^{\mu_i}$  of such type representations will be unitary representations of  $\mathbb{A}$  in separable Hilbert spaces.

**Lemma 1 (Wiener Lemma)** *If  $F \subset L^2(\widehat{\mathbb{A}}, \mu)$  is a closed  $V_a^\mu$ -invariant subspace for all  $a \in \mathbb{A}$  then  $F = 1_Y L^2(\widehat{\mathbb{A}}, \mathcal{B}(\widehat{\mathbb{A}}), \mu)$  for some Borel subset  $Y \subset \widehat{\mathbb{A}}$ .*

Notice however that since  $\mathbb{A}$  is not compact (equivalently,  $\widehat{\mathbb{A}}$  is not discrete), we can find  $\mu$  continuous and therefore  $V^\mu$  has no irreducible (non-zero) subrepresentation. From now on only unitary representations of  $\mathbb{A}$  in separable Hilbert spaces will be considered and we will show how to classify them.

A function  $r: \mathbb{A} \rightarrow \mathbb{C}$  is called *positive definite* if

$$\sum_{n,m=0}^N r(a_n - a_m) z_n \overline{z_m} \geq 0 \quad (1)$$

for each  $N > 0$ ,  $(a_n) \subset \mathbb{A}$  and  $(z_n) \subset \mathbb{C}$ . The central result about positive definite functions is the following theorem (see e. g. [173]).

**Theorem 1 (Bochner–Herglotz)** *Let  $r: \mathbb{A} \rightarrow \mathbb{C}$  be continuous. Then  $r$  is positive definite if and only if there exists (a unique)  $\sigma \in M^+(\widehat{\mathbb{A}})$  such that*

$$r(a) = \int_{\widehat{\mathbb{A}}} \chi(a) d\sigma(\chi) \quad \text{for each } a \in \mathbb{A}.$$

If now  $\mathcal{U} = (U_a)_{a \in \mathbb{A}}$  is a representation of  $\mathbb{A}$  in  $H$  then for each  $x \in H$  the function  $r(a) := \langle U_a x, x \rangle$  is continuous and satisfies (1), so it is positive definite. By the Bochner–Herglotz Theorem there exists a unique measure  $\sigma_{\mathcal{U}, x} = \sigma_x \in M^+(\widehat{\mathbb{A}})$  (called the *Spectral measure* of  $x$ ) such that

$$\widehat{\sigma}_x(a) := \int_{\widehat{\mathbb{A}}} i(a)(\chi) d\sigma_x(\chi) = \langle U_a x, x \rangle$$

for each  $a \in \mathbb{A}$ . Since the partial map  $U_a x \mapsto i(a) \in L^2(\widehat{\mathbb{A}}, \sigma_x)$  is isometric and equivariant, there exists a unique extension of it to a unitary operator

$W: \mathbb{A}(x) \rightarrow L^2(\widehat{\mathbb{A}}, \sigma_x)$  giving rise to an isomorphism of  $\mathcal{U}|_{\mathbb{A}(x)}$  and  $V^{\sigma_x}$ . Then the existence of a spectral decomposition is proved by making use of separability and a choice of maximal cyclic spaces at every step of an induction procedure. Moreover, a spectral decomposition is unique in the following sense.

**Theorem 2 (Spectral Theorem)** *If  $H = \bigoplus_{i=1}^{\infty} \mathbb{A}(x_i) = \bigoplus_{i=1}^{\infty} \mathbb{A}(x'_i)$  are two spectral decompositions of  $H$  then  $\sigma_{x_i} \equiv \sigma_{x'_i}$  for each  $i \geq 1$ .*

It follows that the representation  $\mathcal{U}$  is entirely determined by types (the sets of equivalent measures to a given one) of a decreasing sequence of measures or, equivalently,  $\mathcal{U}$  is determined by its maximal spectral type  $\sigma_{\mathcal{U}}$  and its multiplicity function  $M_{\mathcal{U}}$ .

Notice that eigenvalues of  $\mathcal{U}$  correspond to Dirac measures:  $\chi \in \widehat{\mathbb{A}}$  is an eigenvalue (i. e. for some  $\|x\| = 1$ ,  $U_a(x) = \chi(a)x$  for each  $a \in \mathbb{A}$ ) if and only if  $\sigma_{\mathcal{U},x} = \delta_{\chi}$ . Therefore  $\mathcal{U}$  has a discrete spectrum if and only if the maximal spectral type of  $\mathcal{U}$  is a discrete measure.

We refer the reader to [64,105,127,147,156] for presentations of spectral theory needed in the theory of dynamical systems – such presentations are usually given for  $\mathbb{A} = \mathbb{Z}$  but once we have the Bochner–Herglotz Theorem and the Wiener Lemma, their extensions to the general case are straightforward.

### Koopman Representations

We will consider *measure-preserving* representations of  $\mathbb{A}$ . It means that we fix a probability standard Borel space  $(X, \mathcal{B}, \mu)$  and by  $\text{Aut}(X, \mathcal{B}, \mu)$  we denote the group of automorphisms of this space, that is  $T \in \text{Aut}(X, \mathcal{B}, \mu)$  if  $T: X \rightarrow X$  is a bimeasurable (a.e.) bijection satisfying  $\mu(A) = \mu(TA) = \mu(T^{-1}A)$  for each  $A \in \mathcal{B}$ . Consider then a representation of  $\mathbb{A}$  in  $\text{Aut}(X, \mathcal{B}, \mu)$  that is a group homomorphism  $a \mapsto T_a \in \text{Aut}(X, \mathcal{B}, \mu)$ ; we write  $\mathcal{T} = (T_a)_{a \in \mathbb{A}}$ . Moreover, we require that the associated Koopman representation  $\mathcal{U}_{\mathcal{T}}$  is continuous. Unless explicitly stated,  $\mathbb{A}$ -actions are assumed to be *free*, that is we assume that for  $\mu$ -a.e.  $x \in X$  the map  $a \mapsto T_a x$  is 1–1. In fact, since constant functions are obviously invariant for  $U_{T_a}$ , that is the trivial character 1 is always an eigenvalue of  $\mathcal{U}_{\mathcal{T}}$ , the Koopman representation is considered only on the subspace  $L_0^2(X, \mathcal{B}, \mu)$  of zero mean functions. We will restrict our attention only to *ergodic* dynamical systems (see the glossary). It is easy to see that  $\mathcal{T}$  is ergodic if and only if whenever  $A \in \mathcal{B}$  and  $A = T_a(A)$  ( $\mu$ -a.e.) for all  $a \in \mathbb{A}$  then  $\mu(A)$  equals 0 or 1. In case of ergodic Koopman representations, all eigenvalues are simple. In particular, (ergodic)

Koopman representations with discrete spectra have simple spectra. The reader is referred to monographs mentioned above as well as to [26,158,177,196,204] for basic facts on the theory of dynamical systems.

The passage  $\mathcal{T} \mapsto \mathcal{U}_{\mathcal{T}}$  can be seen as functorial (contravariant). In particular a measure-theoretic isomorphism of  $\mathbb{A}$ -systems  $\mathcal{T}$  and  $\mathcal{T}'$  implies spectral isomorphism of the corresponding Koopman representations; hence spectral properties are measure-theoretic invariants. Since unitary representations are completely classified, one of the main questions in the spectral theory of dynamical systems is to decide which pairs  $([\sigma], M)$  can be realized by Koopman representations. The most spectacular, still unsolved, is the Banach problem concerning  $([\lambda_{\mathbb{T}}], M \equiv 1)$ . Another important problem is to give complete spectral classification in some classes of dynamical systems (classically, it was done in the theory of Kolmogorov and Gaussian dynamical systems). We will also see how spectral properties of dynamical systems can determine their statistical (ergodic) properties; a culmination given by the fact that a spectral isomorphism may imply measure-theoretic similitude (discrete spectrum case, Gaussian–Kronecker case). We conjecture that a dynamical system  $\mathcal{T}$  either is spectrally determined or there are uncountably many pairwise non-isomorphic systems spectrally isomorphic to  $\mathcal{T}$ .

We could also consider Koopman representations in  $L^p$  for  $1 \leq p \neq 2$ . However whenever  $W: L^p(X, \mathcal{B}, \mu) \rightarrow L^p(Y, \mathcal{C}, \nu)$  is a surjective isometry and  $W \circ U_{T_a} = U_{S_a} \circ W$  for each  $a \in \mathbb{A}$  then by the Lamperti Theorem (e. g. [172]) the isometry  $W$  has to come from a non-singular pointwise map  $R: Y \rightarrow X$  and, by ergodicity,  $R$  “preserves” the measure  $\nu$  and hence establishes a measure-theoretic isomorphism [94] (see also [127]). Thus spectral classification of such Koopman representations goes back to the measure-theoretic classification of dynamical systems, so it looks hardly interesting. This does not mean that there are no interesting dynamical questions for  $p \neq 2$ . Let us mention still open Thouvenot’s question (formulated in the 1980s) for  $\mathbb{Z}$ -actions: *For every ergodic  $T$  acting on  $(X, \mathcal{B}, \leq)$ , does there exist  $f \in L^1(X, \mathcal{B}, \mu)$  such that the closed linear span of  $f \circ T^n$ ,  $n \in \mathbb{Z}$  equals  $L^1(X, \mathcal{B}, \mu)$ ?*

Iwanik [79,80] proved that if  $T$  is a system with positive entropy then its  $L^p$ -multiplicity is  $\infty$  for all  $p > 1$ . Moreover, Iwanik and de Sam Lazaro [85] proved that for Gaussian systems (they will be considered in Sect. “Spectral Theory of Dynamical Systems of Probabilistic Origin”) the  $L^p$ -multiplicities are the same for all  $p > 1$  (see also [137]).

### Markov Operators, Joinings and Koopman Representations, Disjointness and Spectral Disjointness, Entropy

We would like to emphasize that spectral theory is closely related to the theory of joinings (see ► [Joinings in Ergodic Theory](#) for needed definitions). The elements  $\rho$  of the set  $J(S, \mathcal{T})$  of joinings of two  $\mathbb{A}$ -actions  $S$  and  $\mathcal{T}$  are in a 1-1 correspondence with Markov operators  $J = J_\rho$  between the  $L^2$ -spaces equivariant with the corresponding Koopman representations (see the glossary and ► [Joinings in Ergodic Theory](#)). The set of ergodic self-joinings of an ergodic  $\mathbb{A}$ -action  $\mathcal{T}$  is denoted by  $J_2^e(\mathcal{T})$ .

Each Koopman representation  $\mathcal{U}_{\mathcal{T}}$  consists of Markov operators (indeed,  $U_{T_a}$  is clearly a Markov operator). In fact, if  $U \in U(L^2(X, \mathcal{B}, \mu))$  is Markov then it is of the form  $U_R$ , where  $R \in \text{Aut}(X, \mathcal{B}, \mu)$  [133]. This allows us to see Koopman representations as unitary Markov representations, but since a spectral isomorphism does not “preserve” the set of Markov operators, spectrally isomorphic systems can have drastically different sets of self-joinings.

We will touch here only some aspects of interactions (clearly, far from completeness) between the spectral theory and the theory of joinings.

In order to see however an example of such interactions let us recall that the simplicity of eigenvalues for ergodic systems yields a short “joining” proof of the classical isomorphism theorem of Halmos-von Neumann in the discrete spectrum case: Assume that  $S = (S_{-i})_{i \in \mathbb{A}}$  and  $\mathcal{T} = (T_a)_{a \in \mathbb{A}}$  are ergodic  $\mathbb{A}$ -actions on  $(X, \mathcal{B}, \mu)$  and  $(Y, \mathcal{C}, \nu)$  respectively. Assume that both Koopman representations have purely discrete spectrum and that their group of eigenvalues are the same. Then  $S$  and  $\mathcal{T}$  are measure-theoretically isomorphic. Indeed, each ergodic joining of  $\mathcal{T}$  and  $S$  is the graph of an isomorphism of these two systems (see [127]; see also Goodson’s proof in [66]). Another example of such interactions appear in the study of Rokhlin’s multiple mixing problem and its relation with the pairwise independence property (PID) for joinings of higher order. We will not deal with this subject here, referring the reader to ► [Joinings in Ergodic Theory](#) (see however Sect. “Lifting Mixing Properties”).

Following [60], two  $\mathbb{A}$ -actions  $S$  and  $\mathcal{T}$  are called *disjoint* provided the product measure is the only element in  $J(S, \mathcal{T})$ . It was already noticed in [72] that spectrally disjoint systems are disjoint in the Furstenberg sense; indeed,  $\text{Im}(J_\rho|_{L_0^2}) = \{0\}$  since  $\sigma_{\mathcal{T}, J_\rho f} \ll \sigma_{S, f}$ .

Notice that whenever we take  $\rho \in J_2^e(\mathcal{T})$  we obtain a new ergodic  $\mathbb{A}$ -action  $(T_a \times T_a)_{a \in \mathbb{A}}$  defined on the probability space  $(X \times X, \rho)$ . One can now ask how close spectrally to  $\mathcal{T}$  is this new action? It turns out that ex-

cept of the obvious fact that the marginal  $\sigma$ -algebras are factors,  $(\mathcal{T} \times \mathcal{T}, \rho)$  can have other factors spectrally disjoint with  $\mathcal{T}$ : the most striking phenomenon here is a result of Smorodinsky and Thouvenot [198] (see also [29]) saying that each zero entropy system is a factor of an ergodic self-joining system of a fixed Bernoulli system (Bernoulli systems themselves have countable Haar spectrum). The situation changes if  $\rho = \mu \otimes \mu$ . In this case for  $f, g \in L^2(X, \mathcal{B}, \mu)$  the spectral measure of  $f \otimes g$  is equal to  $\sigma_{\mathcal{T}, f} * \sigma_{\mathcal{T}, g}$ . A consequence of this observation is that an ergodic action  $\mathcal{T} = (T_a)_{a \in \mathbb{A}}$  is weakly mixing (see the glossary) if and only if the product measure  $\mu \otimes \mu$  is an ergodic self-joining of  $\mathcal{T}$ .

The entropy which is a basic measure-theoretic invariant does not appear when we deal with spectral properties. We will not give here any formal definition of entropy for amenable group actions referring the reader to [153]. Assume that  $\mathbb{A}$  is countable and discrete. We always assume that  $\mathbb{A}$  is Abelian, hence it is amenable. For each dynamical system  $\mathcal{T} = (T_a)_{a \in \mathbb{A}}$  acting on  $(X, \mathcal{B}, \mu)$ , we can find a largest invariant sub- $\sigma$  field  $\mathcal{P} \subset \mathcal{B}$ , called the *Pinsker  $\sigma$ -algebra*, such that the entropy of the corresponding quotient system is zero. Generalizing the classical Rokhlin-Sinai Theorem (see also [97] for  $\mathbb{Z}^d$ -actions), Thouvenot (unpublished) and independently Dooley and Golodets [31] proved this theorem for groups even more general than those considered here: *The spectrum of  $\mathcal{U}_{\mathcal{T}}$  on  $L^2(X, \mathcal{B}, \mu) \ominus L^2(\mathcal{P})$  is Haar with uniform infinite multiplicity.* This general result is quite intricate and based on methods introduced to entropy theory by Rudolph and Weiss [179] with a very surprising use of Dye’s Theorem on orbital equivalence of all ergodic systems. For  $\mathbb{A}$  which is not countable the same result was recently proved in [17] in case of unimodular amenable groups which are not increasing union of compact subgroups. It follows that spectral theory of dynamical systems essentially reduces to the zero entropy case.

### Maximal Spectral Type of a Koopman Representation, Alexeyev’s Theorem

Only few general properties of maximal spectral types of Koopman representations are known. The fact that a Koopman representation preserves the space of real functions implies that its maximal spectral type is the type of a symmetric (invariant under the map  $\chi \mapsto \bar{\chi}$ ) measure.

Recall that the *Gelfand spectrum*  $\sigma(\mathcal{U})$  of a representation  $\mathcal{U} = (U_a)_{a \in \mathbb{A}}$  is defined as the set of *approximative eigenvalues* of  $\mathcal{U}$ , that is  $\sigma(\mathcal{U}) \ni \chi \in \hat{\mathbb{A}}$  if for a sequence  $(x_n)$  bounded and bounded away from zero,  $\|U_a x_n - \chi(a) x_n\| \rightarrow 0$  for each  $a \in \mathbb{A}$ . The spectrum



is a closed subset in the topology of pointwise convergence, hence in the compact-open topology of  $\widehat{\mathbb{A}}$ . In case of  $\mathbb{A} = \mathbb{Z}$ , the above set  $\sigma(U)$  is equal to  $\{z \in \mathbb{C} : U - z \cdot Id \text{ is not invertible}\}$ .

Assume now that  $\mathbb{A}$  is countable and discrete (and Abelian). Then there exists a Følner sequence  $(B_n)_{n \geq 1}$  whose elements tile  $\mathbb{A}$  [153]. Take a free and ergodic action  $\mathcal{T} = (T_a)_{a \in \mathbb{A}}$  on  $(X, \mathcal{B}, \mu)$ . By [153] for each  $\varepsilon > 0$  we can find a set  $Y_n \in \mathcal{B}$  such that the sets  $T_b Y_n$  are pairwise disjoint for  $b \in B_n$  and  $\mu(\bigcup_{b \in B_n} T_b Y_n) > 1 - \varepsilon$ . For each  $\chi \in \widehat{\mathbb{A}}$ , by considering functions of the form  $f_n = \sum_{b \in B_n} \chi(b) 1_{T_b Y_n}$  we obtain that  $\chi \in \sigma(\mathcal{U}_{\mathcal{T}})$ . It follows that the topological support of the maximal spectral type of the Koopman representation of a free and ergodic action is full [105, 127, 147]. The theory of Gaussian systems shows in particular that there are symmetric measures on the circle whose topological support is the whole circle but which cannot be maximal spectral types of Koopman representations.

An open well-known question remains whether an absolutely continuous measure  $\rho$  is the maximal spectral type of a Koopman representation if and only if  $\rho$  is equivalent to a Haar measure of  $\widehat{\mathbb{A}}$  (this is unknown for  $\mathbb{A} = \mathbb{Z}$ ).

Another interesting question was recently raised by A. Katok (private communication): *Is it true that the topological supports of all measures in a spectral sequence of a Koopman representation are full?* If the answer to this question is positive then for example the essential supremum of  $M_{\mathcal{U}_{\mathcal{T}}}$  is the same on all balls of  $\widehat{\mathbb{A}}$ .

Notice that the fact that all spectral measures in a spectral sequence are symmetric means that  $\mathcal{U}_{\mathcal{T}}$  is isomorphic to  $\mathcal{U}_{\mathcal{T}^{-1}}$ . A. del Junco [89] showed that generically for  $\mathbb{Z}$ -actions,  $T$  and its inverse are not measure-theoretically isomorphic (in fact he proved disjointness).

Let  $\mathcal{T}$  be an  $\mathbb{A}$ -action on  $(X, \mathcal{B}, \mu)$ . One can ask whether a “good” function can realize the maximal spectral type of  $\mathcal{U}_{\mathcal{T}}$ . In particular can we find a function  $f \in L^\infty(X, \mathcal{B}, \mu)$  that realizes the maximal spectral type? The answer is given in the following general theorem (see [139]).

**Theorem 3 (Alexeyev’s Theorem)** *Assume that  $\mathcal{U} = (U_a)_{a \in \mathbb{A}}$  is a unitary representation of  $\mathbb{A}$  in a separable Hilbert space  $H$ . Assume that  $F \subset H$  is a dense linear subspace. Assume moreover that with some  $F$ -norm  $|\cdot|$  – stronger than the norm  $\|\cdot\|$  given by the scalar product –  $F$  becomes a Fréchet space. Then, for each spectral measure  $\sigma (\ll \sigma_{\mathcal{U}})$  there exists  $y \in F$  such that  $\sigma_y \gg \sigma$ . In particular, there exists  $y \in F$  that realizes the maximal spectral type.*

By taking  $H = L^2(X, \mathcal{B}, \mu)$  and  $F = L^\infty(X, \mathcal{B}, \mu)$  we obtain the positive answer to the original question. Alexeyev [14] proved the above theorem for  $\mathbb{A} = \mathbb{Z}$  using

analytic functions. Refining Alexeyev’s original proof, Frączek [52] showed the existence of a sufficiently regular function realizing the maximal spectral type depending only on the “regularity” of the underlying probability space, e.g. when  $X$  is a compact metric space (compact manifold) then one can find a continuous (smooth) function realizing the maximal spectral type.

By the theory of systems of probabilistic origin (see Sect. “Spectral Theory of Dynamical Systems of Probabilistic Origin”), in case of simplicity of the spectrum, one can easily point out spectral measures whose types are not realized by (essentially) bounded functions. However, it is still an open question whether for each Koopman representation  $\mathcal{U}_{\mathcal{T}}$  there exists a sequence  $(f_i)_{i \geq 1} \subset L^\infty(X, \mathcal{B}, \mu)$  such that the sequence  $(\sigma_{f_i})_{i \geq 1}$  is a spectral sequence for  $\mathcal{U}_{\mathcal{T}}$ . For  $\mathbb{A} = \mathbb{Z}$  it is unknown whether the maximal spectral type of a Koopman representation can be realized by a characteristic function.

## Spectral Theory of Weighted Operators

We now pass to the problem of possible essential values for the multiplicity function of a Koopman representation. However, one of known techniques is a use of cocycles, so before we tackle the multiplicity problem, we will go through recent results concerning spectral theory of compact group extensions automorphisms which in turn entail a study of weighted operators (see the glossary).

Assume that  $T$  is an ergodic automorphism of a standard Borel probability space  $(X, \mathcal{B}, \mu)$ . Let  $\xi: X \rightarrow \mathbb{T}$  be a measurable function and let  $V = V_{\xi, T}$  be the corresponding weighted operator. To see a connection of weighted operators with Koopman representations of compact group extensions consider a compact (metric) Abelian group  $G$  and a cocycle  $\varphi: X \rightarrow G$ . Then  $U_{T_\varphi}$  (see the glossary) acts on  $L^2(X \times G, \mu \otimes \lambda_G)$ . But

$$L^2(X \times G, \mu \otimes \lambda_G) = \bigoplus_{\chi \in \widehat{G}} L_\chi, \quad \text{where } L_\chi = L^2(X, \mu) \otimes \chi,$$

where  $L_\chi$  is a  $U_{T_\varphi}$ -invariant (clearly, closed) subspace. Moreover, the map  $f \otimes \chi \mapsto f$  settles a unitary isomorphism of  $U_{T_\varphi}|_{L_\chi}$  with the operator  $V_{\chi \circ \varphi, T}$ . Therefore, spectral analysis of such Koopman representations reduces to the spectral analysis of weighted operators  $V_{\chi \circ \varphi, T}$  for all  $\chi \in \widehat{G}$ .

## Maximal Spectral Type of Weighted Operators over Rotations

The spectral analysis of weighted operators  $V_{\xi, T}$  is especially well developed in case of rotations, i.e. when



we additionally assume that  $T$  is an ergodic rotation on a compact monothetic group  $X$ :  $Tx = x + x_0$ , where  $x_0$  is a topologically cyclic element of  $X$  (and  $\mu$  will stand for Haar measure  $\lambda_x$  of  $X$ ). In this case Helson's analysis [74] applies (see also [68,82,127,160]) leading to the following conclusions:

- The maximal spectral type  $\sigma_{V_{\xi,T}}$  is either discrete or continuous.
- When  $\sigma_{V_{\xi,T}}$  is continuous it is either singular or Lebesgue.
- The spectral multiplicity of  $V_{\xi,T}$  is uniform.

We now pass to a description of some results in case when  $Tx = x + \alpha$  is an irrational rotation on the additive circle  $X = [0, 1)$ . It was already noticed in the original paper by Anzai [16] that when  $\xi: X \rightarrow \mathbb{T}$  is an affine cocycle ( $\xi(x) = \exp(nx + c)$ ,  $0 \neq n \in \mathbb{Z}$ ) then  $V_{\xi,T}$  has a Lebesgue spectrum. It was then considered by several authors (originated by [123], see also [24,26]) to which extent this property is stable when we perturb our cocycle. Since the topological degree of affine cocycles is different from zero, when perturbing them we consider smooth perturbations by cocycles of degree zero.

**Theorem 4 ([82])** *Assume that  $Tx = x + \alpha$  is an irrational rotation. If  $\xi: [0, 1) \rightarrow \mathbb{T}$  is of non-zero degree, absolutely continuous, with the derivative of bounded variation then  $V_{\xi,T}$  has a Lebesgue spectrum.*

In the same paper, it is noticed that if we drop the assumption on the derivative then the maximal spectral type of  $V_{\xi,T}$  is a Rajchman measure (i. e. its Fourier transform vanishes at infinity). It is still an open question, whether one can find  $\xi$  absolutely continuous with non-zero degree and such that  $V_{\xi,T}$  has singular spectrum. "Below" absolute continuity, topological properties of the cocycle seem to stop playing any role – in [82] a continuous, degree 1 cocycle  $\xi$  of bounded variation is constructed such that  $\xi(x) = \eta(x)/\eta(Tx)$  for a measurable  $\eta: [0, 1) \rightarrow \mathbb{T}$  (that is  $\xi$  is a *coboundary*) and therefore  $V_{\xi,T}$  has purely discrete spectrum (it is isomorphic to  $U_T$ ). For other results about Lebesgue spectrum for Anzai skew products see also [24,53,81] (in [53]  $\mathbb{Z}^d$ -actions of rotations and so called winding numbers instead of topological degree are considered).

When the cocycle is still smooth but its degree is zero the situation drastically changes. Given an absolutely continuous function  $f: [0, 1) \rightarrow \mathbb{R}$  M. Herman [76], using the Denjoy–Koksma inequality (see e. g. [122]), showed that  $f_0^{(q_n)} \rightarrow 0$  uniformly (here  $f_0 = f - \int_0^1 f d\lambda_{[0,1]}$  and  $(q_n)$  stands for the sequence of denominators of  $\alpha$ ). It follows that  $T_{e^{2\pi i f}}$  is rigid and hence has a singular spec-

trum. B. Fayad [37] shows that this result is no longer true if one dimensional rotation is replaced by a multi-dimensional rotation (his counterexample is in the analytic class). See also [130] for the singularity of spectrum for functions  $f$  whose Fourier transform satisfies  $o(1/|n|)$  condition or to [84], where it is shown that sufficiently small variation implies singularity of the spectrum.

A natural class of weighted operators arises when we consider Koopman operators of rotations on nil-manifolds. We only look at the particular example of such a rotation on a quotient of the Heisenberg group  $(\mathbb{R}^3, *)$  (a general spectral theory of nil-actions was mainly developed by W. Parry [157]) – these actions have countable Lebesgue spectrum in the orthocomplement of the subspace of eigenfunctions) that is take the nil-manifold  $\mathbb{R}^3/_{*}\mathbb{Z}^3$  on which we define the nil-rotation  $S((x, y, z) * \mathbb{Z}^3) = (\alpha, \beta, 0) * (x, y, z) * \mathbb{Z}^3 = (x + \alpha, y + \beta, z + \alpha y) * \mathbb{Z}^3$ , where  $\alpha, \beta$  and 1 are rationally independent. It can be shown that  $S$  is isomorphic to the skew product defined on  $[0, 1)^2 \times \mathbb{T}$  by

$$T_{\varphi}: (x, y, z) \mapsto (x + \alpha, y + \beta, z \cdot e^{2\pi i \varphi(x, y)}),$$

where  $\varphi(x, y) = \alpha y - \psi(x + \alpha, y + \beta) + \psi(x, y)$  with  $\psi(x, y) = x[y]$ . Since nil-cocycles can be considered as a certain analog of affine cocycles for one-dimensional rotations, it would be nice to explain to what kind of perturbations the Lebesgue spectrum property is stable.

Yet another interesting problem which is related to the spectral theory of extensions given by so called *Rokhlin cocycles* (see Sect. "Rokhlin Cocycles") arises, when given  $f: [0, 1) \rightarrow \mathbb{R}$ , we want to describe spectrally the one-parameter set of weighted operators  $W_c := V_{e^{2\pi i c f}, T}$ ; here  $T$  is a fixed irrational rotation by  $\alpha$ . As proved by quite sophisticated arguments in [84], if we take  $f(x) = x$  then for all non-integer  $c \in \mathbb{R}$  the spectrum of  $W_c$  is continuous and singular (see also [68] and [145] where some special  $\alpha$ 's are considered). It has been open for some time if at all one can find  $f: [0, 1) \rightarrow \mathbb{R}$  such that for each  $c \neq 0$ , the operator  $W_c$  has a Lebesgue spectrum. The positive answer is given in [205]: for example if  $f(x) = x^{-(2+\varepsilon)}$  ( $\varepsilon > 0$ ) and  $\alpha$  has bounded partial quotients then  $W_c$  has a Lebesgue spectrum for all  $c \neq 0$ . All functions with such a property considered in [205] are non-integrable. It would be interesting to find an integrable  $f$  with the above property.

We refer to [66] and the references therein for further results especially for transformations of the form  $(x, y) \mapsto (x + \alpha, 1_{[0, \beta)}(x) + y)$  on  $[0, 1) \times \mathbb{Z}/2\mathbb{Z}$ . Recall however that earlier Katok and Stepin [104] used this kind of transfor-

mations to give a first counterexample to the Kolmogorov group property (see the glossary) for the spectrum.

### The Multiplicity Problem for Weighted Operators over Rotations

In case of perturbations of affine cocycles, this problem was already raised by Kushnirenko [123]. Some significant results were obtained by M. Guenais. Before we state her results let us recall a useful criterion to find an upper bound for the multiplicity: *If there exist  $c > 0$  and a sequence  $(F_n)_{n \geq 1}$  of cyclic subspaces of  $H$  such that for each  $y \in H$ ,  $\|y\| = 1$  we have  $\liminf_{n \rightarrow \infty} \|\text{proj}_{F_n} y\|^2 \geq c$ , then  $\text{esssup}(M_U) \leq 1/c$  which follows from a well-known lemma of Chacon [23, 26, 111, 127]. Using this and a technique which is close to the idea of local rank one (see [44, 111]) M. Guenais [69] proved a series of results on multiplicity which we now list.*

**Theorem 5** *Assume that  $Tx = x + \alpha$  and let  $\xi: [0, 1) \rightarrow \mathbb{T}$  be a cocycle.*

- (i) *If  $\xi(x) = e^{2\pi i c x}$  then  $M_{V_{\xi, T}}$  is bounded by  $|c| + 1$ .*
- (ii) *If  $\xi$  is absolutely continuous and  $\xi$  is of topological degree zero, then  $V_{\xi, T}$  has a simple spectrum.*
- (iii) *if  $\xi$  is of bounded variation, then*  

$$M_{V_{\xi, T}} \leq \max(2, 2\pi \text{Var}(\xi)/3).$$

### Remarks on the Banach Problem

We already mentioned in Introduction the Banach problem in ergodic theory, which is simply the question whether there exists a Koopman representation for  $\mathbb{A} = \mathbb{Z}$  with simple Lebesgue spectrum. In fact no example of a Koopman representation with Lebesgue spectrum of finite multiplicity is known. Helson and Parry [75] asked for the validity of a still weaker version: *Can one construct  $T$  such that  $U_T$  has a Lebesgue component in its spectrum whose multiplicity is finite?* Quite surprisingly in [75] they give a general construction yielding for each ergodic  $T$  a cocycle  $\varphi: X \rightarrow \mathbb{Z}/2\mathbb{Z}$  such that the unitary operator  $U_{T_\varphi}$  has a Lebesgue spectrum in the orthocomplement of functions depending only on the  $X$ -coordinate. Then Mathew and Nadkarni [144] gave examples of cocycles over so called dyadic adding machine for which the multiplicity of the Lebesgue component was equal to 2. In [126] this was generalized to so called Toeplitz  $\mathbb{Z}/2\mathbb{Z}$ -extensions of adding machines: for each even number  $k$  we can find a two-point extension of an adding machine so that the multiplicity of the Lebesgue component is  $k$ . Moreover, it was shown that Mathew and Nadkarni's constructions from [144] in fact are close to sys-

tems arising from number theory (like the famous Rudin-Shapiro sequence, e.g. [160]), relating the result about multiplicity of the Lebesgue component to results by Kamae [96] and Queffelec [160]. Independently of [126], Ageev [8] showed that one can construct 2-point extensions of the Chacon transformation realizing (by taking powers of the extension) each even number as the multiplicity of the Lebesgue component. Contrary to all previous examples, Ageev's constructions are weakly mixing.

Still an open question remains whether for  $\mathbb{A} = \mathbb{Z}$  one can find a Koopman representation with the Lebesgue component of multiplicity 1 (or even odd).

In [70], M. Guenais studies the problem of Lebesgue spectrum in the classical case of Morse sequences (see [107] as well as [124], where the problem of spectral classification in this class is studied). All dynamical systems arising from Morse sequences have simple spectra [124]. It follows that if a Lebesgue component appears in a Morse dynamical system, it has multiplicity one. Guenais [70] using a Riesz product technique relates the Lebesgue spectrum problem with the still open problem of whether a construction of "flat" trigonometric polynomials with coefficients  $\pm 1$  is possible. However, it is proved in [70] that such a construction can be carried out on some compact Abelian groups and it leads, for an Abelian countable torsion group  $\mathbb{A}$ , to a construction of an ergodic action of  $\mathbb{A}$  with simple spectrum and a Haar component in its spectrum.

### Lifting Mixing Properties

We now give one more example of interactions between spectral theory and joinings (see Introduction) that gives rise to a quick proof of the fact that  $r$ -fold mixing property of  $T$  ( $r \geq 2$ ) lifts to a weakly mixing compact group extension  $T_\varphi$  (the original proof of this fact is due to D. Rudolph [175]). Indeed, to prove  $r$ -fold mixing for a mixing (= 2-mixing) transformation  $S$  (acting on  $(Y, \mathcal{C}, \nu)$ ) one has to prove that each weak limit of off-diagonal self-joinings (given by powers of  $S$ , see ► [Joinings in Ergodic Theory](#)) of order  $r$  is simply the product measure  $\nu^{\otimes r}$ . We need also a Furstenberg's lemma [62] about relative unique ergodicity (RUE) of compact group extensions  $T_\varphi$ : *If  $\mu \otimes \lambda_G$  is an ergodic measure for  $T_\varphi$  then it is the only (ergodic) invariant measure for  $T_\varphi$  whose projection on the first coordinate is  $\mu$ .* Now the result about lifting  $r$ -fold mixing to compact group extensions follows directly from the fact that whenever  $T_\varphi$  is weakly mixing,  $(\mu \otimes \lambda_G)^{\otimes r}$  is an ergodic measure (this approach was shown to me by A. del Junco). In particular if  $T$  is mixing and  $T_\varphi$  is weakly

mixing then for each  $\chi \in \widehat{G} \setminus \{1\}$ , the maximal spectral type of  $V_{\chi \circ \varphi, T}$  is Rajchman.

See Sect. “Rokhlin Cocycles” for a generalization of the lifting result to Rokhlin cocycle extensions.

### The Multiplicity Function

In this chapter only  $\mathbb{A} = \mathbb{Z}$  is considered (for other groups, even for  $\mathbb{R}$ , much less is known; see however the case of so called *product  $\mathbb{Z}^d$ -actions* [50]). Contrary to the case of maximal spectral type, it is rather commonly believed that there are no restrictions for the set of essential values of Koopman representations.

### Cocycle Approach

We will only concentrate on some results of the last twenty years. In 1983, E.A. Robinson [164] proved that for each  $n \geq 1$  there exists an ergodic transformation whose maximal spectral multiplicity is  $n$ . Another important result was proved in [165] (see also [98]), where it is shown that given a finite set  $M \subset \mathbb{N}$  containing 1 and closed under the least common multiple one can find (even a weakly mixing)  $T$  so that the set of essential values of the multiplicity function equals  $M$ . This result was then extended in [67] to infinite sets and finally in [125] (see also [11]) to all subsets  $M \subset \mathbb{N}$  containing 1. In fact, as we have already noticed in the previous section the spectral theory for compact Abelian group extensions is reduced to a study of weighted operators and then to comparing maximal spectral types for such operators. This leads to sets of the form

$$M(G, \nu, H) = \left\{ \#(\{\chi \circ \nu^i : i \in \mathbb{Z}\} \cap \text{an}ih(H)) : \chi \in \text{an}ih(H) \right\}$$

( $H \subset G$  is a closed subgroup and  $\nu$  is a continuous group automorphism of  $G$ ). Then an algebraic lemma has been proved in [125] saying that each set  $M$  containing 1 is of the form  $M(G, \nu, H)$  and the techniques to construct “good” cocycles and a passage to “natural factors” yielded the following: *For each  $M \subset \{1, 2, \dots\} \subset \{\infty\}$  containing 1 there exists an ergodic automorphism such that the set of essential values for its Koopman representation equals  $M$ .* See also [166] for the case of non-Abelian group extensions.

A similar in spirit approach (that means, a passage to a family of factors) is present in a recent paper of Ageev [13] in which he first applies Katok’s analysis (see [98, 102]) for spectral multiplicities of the Koopman representation associated with Cartesian products  $T^{\times k}$  for

a generic transformation  $T$ . In a natural way this approach leads to study multiplicities of tensor products of unitary operators. Roughly, the multiplicity is computed as the number of atoms (counted modulo obvious symmetries) for conditional measures (see [98]) of a product measure over its convolution. Ageev [13] proved that for a typical automorphism  $T$  the set of the values of the multiplicity function for  $U_{T^{\times k}}$  equals  $\{k, k(k-1), \dots, k!\}$  and then he just passes to “natural” factors for the Cartesian products by taking sets invariant under a fixed subgroup of permutations of coordinates. In particular, he obtains all sets of the form  $\{2, 3, \dots, n\}$  on  $L_0^2$ . He also shows that such sets of multiplicities are realizable in the category of mixing transformations.

### Rokhlin’s Uniform Multiplicity Problem

The Rokhlin multiplicity problem (see the recent book by Anosov [15]) was, given  $n \geq 2$ , to construct an ergodic transformation with uniform multiplicity  $n$  on  $L_0^2$ . A solution for  $n = 2$  was independently given by Ageev [9] and Ryzhikov [188] (see also [15] and [66]) and in fact it consists in showing that for some  $T$  (actually, any  $T$  with simple spectrum for which  $1/2(Id + U_T)$  is in the weak operator closure of the powers of  $U_T$  will do) the multiplicity of  $T \times T$  is uniformly equal to 2 (see also Sect. “Future Directions”).

In [12], Ageev proposed a new approach which consists in considering actions of “slightly non-Abelian” groups; and showing that for a “typical” action of such a group a fixed “direction” automorphism has a uniform multiplicity. Shortly after publication of [12], Danilenko [27], following Ageev’s approach, considerably simplified the original proof. We will present Danilenko’s arguments.

Fix  $n \geq 1$ . Denote  $\bar{e}_i = (0, \dots, 1, \dots, 0) \in \mathbb{Z}^n$ ,  $i = 1, \dots, n$ . We define an automorphism  $L$  of  $\mathbb{Z}^n$  setting  $L(\bar{e}_i) = \bar{e}_{i+1}$ ,  $i = 1, \dots, n-1$  and  $L(\bar{e}_n) = \bar{e}_1$ . Using  $L$  we define a semi-direct product  $G := \mathbb{Z}^n \rtimes \mathbb{Z}$  defining multiplication as  $(u, k) \cdot (w, l) = (u + L^k w, k + l)$ . Put  $e_0 = (0, 1)$ ,  $e_i = (\bar{e}_i, 0)$ ,  $i = 1, \dots, n$  (and  $Le_i = (L\bar{e}_i, 0)$ ). Moreover, denote  $e_{n+1} = e_0^n = (0, n)$ . Notice that  $e_0 \cdot e_i \cdot e_0^{-1} = Le_i$  for  $i = 1, \dots, n$  ( $L(e_{n+1}) = e_{n+1}$ ).

**Theorem 6 (Ageev, Danilenko)** *For every unitary representation  $\mathcal{U}$  of  $G$  in a separable Hilbert space  $H$ , for which  $U_{e_1 - L^r e_1}$  has no non-trivial fixed points for  $1 \leq r < n$ , the essential values of the multiplicity function for  $U_{e_{n+1}}$  are contained in the set of multiples of  $n$ . If, in addition,  $U_{e_0}$  has a simple spectrum, then  $U_{e_{n+1}}$  has uniform multiplicity  $n$ .*

It is then a certain work to show that the assumption of the second part of the theorem is satisfied for a typical action of the group  $G$ . Using a special  $(C, F)$ -construction with all the cut-and-stack parameters explicit Danilenko [27] was also able to show that each set of the form  $k \cdot M$ , where  $k \geq 1$  and  $M$  is an arbitrary subset of natural numbers containing 1, is realizable as the set of essential values of a Koopman representation.

Some other constructions based on the solution of the Rokhlin problem for  $n = 2$  and the method of [125] are presented in [103] leading to sets different than those pointed above; these sets contain 2 as their minimum.

### Rokhlin Cocycles

We consider now a certain class of extensions which should be viewed as a generalization of the concept of compact group extensions. We will focus on  $\mathbb{Z}$ -actions only.

Assume that  $T$  is an ergodic automorphism of  $(X, \mathcal{B}, \mu)$ . Let  $G$  be a l.c.s.c. Abelian group. Assume that this group acts on  $(Y, \mathcal{C}, \nu)$ , that is we have a  $G$ -action  $S = (S_g)_{g \in G}$  on  $(Y, \mathcal{C}, \nu)$ . Let  $\varphi: X \rightarrow G$  be a cocycle. We then define an automorphism  $T_{\varphi, S}$  of the space  $(X \times Y, \mathcal{B} \otimes \mathcal{C}, \mu \otimes \nu)$  by

$$T_{\varphi, S}(x, y) = (Tx, S_{\varphi(x)}(y)).$$

Such an extension is called a *Rokhlin cocycle extension* (the map  $x \mapsto S_{\varphi(x)}$  is called a *Rokhlin cocycle*). Such an operation generalizes the case of compact group extensions; indeed, when  $G$  is compact the action of  $G$  on itself by rotations preserves Haar measure. (It is quite surprising, that when only we admit  $G$  non-Abelian, then, as shown in [28], each ergodic extension of  $T$  has a form of a Rokhlin cocycle extension.) Ergodic and spectral properties of such extensions are examined in several papers: [63, 65, 129, 131, 132, 133, 167, 176]. Since in these papers rather joining aspects are studied (among other things in [129] Furstenberg's RUE lemma is generalized to this new context), we will mention here only few results, mainly spectral, following [129] and [133]. We will constantly assume that  $G$  is non-compact. As  $\varphi: X \rightarrow G$  is then a cocycle with values in a non-compact group, the theory of such cocycles is much more complicated (see e.g. [193]), and in fact the theory of Rokhlin cocycle extensions leads to interesting interactions between classical ergodic theory, the theory of cocycles and the theory of non-singular actions arising from cocycles taking values in non-compact groups – especially, the Mackey action associated to  $\varphi$  plays a crucial role here (see the problem of invariant measures for  $T_{\varphi, S}$  in [132] and [28]);

see also monographs [1, 98, 101, 193]. Especially, two Borel subgroups of  $\widehat{G}$  are important here:

$$\Sigma_{\varphi} = \{\chi \in \widehat{G}: \chi \circ \varphi = c \cdot \xi / \xi \circ T \text{ for a measurable } \xi: X \rightarrow \mathbb{T} \text{ and } c \in \mathbb{T}\}.$$

and its subgroup  $\Lambda_{\varphi}$  given by  $c = 1$ .  $\Lambda_{\varphi}$  turns out to be the group of  $L^{\infty}$ -eigenvalues of the Mackey action (of  $G$ ) associated to the cocycle  $\varphi$ . This action is the quotient action of the natural action of  $G$  (by translations on the second coordinate) on the space of ergodic components of the skew product  $T_{\varphi}$  – the Mackey action is (in general) not measure-preserving, it is however non-singular. We refer the reader to [2, 78, 147] for other properties of those subgroups.

#### Theorem 7 ([132, 133])

- (i)  $\sigma_{T_{\varphi, S}}|_{L^2(X \times Y, \mu \otimes \nu) \ominus L^2(X, \mu)} = \int_{\widehat{G}} \sigma_{V_{\chi \circ \varphi, T}} d\sigma_S$ .
- (ii)  $T_{\varphi, S}$  is ergodic if and only if  $T$  is ergodic and  $\sigma_S(\Lambda_{\varphi}) = 0$ .
- (iii)  $T_{\varphi, S}$  is weakly mixing if and only if  $T$  is weakly mixing and  $S$  has no eigenvalues in  $\Sigma_{\varphi}$ .
- (iv) if  $T$  is mixing,  $S$  is mildly mixing,  $\varphi$  is recurrent and not cohomologous to a cocycle with values in a compact subgroup of  $G$  then  $T_{\varphi, S}$  remains mixing.
- (v) If  $T$  is  $r$ -fold mixing,  $\varphi$  is recurrent and  $T_{\varphi, S}$  is mildly mixing then  $T_{\varphi, S}$  is also  $r$ -fold mixing.
- (vi) If  $T$  and  $R$  are disjoint, the cocycle  $\varphi$  is ergodic and  $S$  is mildly mixing then  $T_{\varphi, S}$  remains disjoint with  $R$ .

Let us recall [61, 195] that an  $\mathbb{A}$ -action  $S = (S_a)_{a \in \mathbb{A}}$  is mildly mixing (see the glossary) if and only if the  $\mathbb{A}$ -action  $(S_a \times \tau_a)_{a \in \mathbb{A}}$  remains ergodic for every properly ergodic non-singular  $\mathbb{A}$ -action  $\tau = (\tau_a)_{a \in \mathbb{A}}$ .

Coming back to Smorodinsky–Thouvenot's result about factors of ergodic self-joinings of a Bernoulli automorphism we would like to emphasize here that the disjointness result (vi) above was used in [132] to give an example of an automorphism which is disjoint from all weakly mixing transformations but which has an ergodic self-joining whose associated automorphism has a non-trivial weakly mixing factor. In a sense this is opposed to Smorodinsky–Thouvenot's result as here from self-joinings we produced a “more complicated” system (namely the weakly mixing factor) than the original system.

It would be interesting to develop the theory of spectral multiplicity for Rokhlin cocycle extensions as it was done in the case of compact group extensions. However a difficulty is that in the compact group extension case we deal with a countable direct sum of representations of the form



$V_{\chi \circ \varphi, T}$  while in the non-compact case we have to consider an integral of such representations.

### Rank-1 and Related Systems

Although properties like mixing, weak (and mild) mixing as well as ergodicity, are clearly spectral properties, they have “good” measure-theoretic formulations (expressed by a certain behavior on sets). Simple spectrum property is another example of a spectral property, and it was a popular question in the 1980s whether simple spectrum property of a Koopman representation can be expressed in a more “measure-theoretic” way. We now recall rank-1 concept which can be seen as a notion close to Katok’s and Stepin’s theory of cyclic approximation [104] (see also [26]).

Assume that  $T$  is an automorphism of a standard probability Borel space  $(X, \mathcal{B}, \mu)$ .  $T$  is said to have *rank one* property if there exists an increasing sequence of Rokhlin towers tending to the partition into points (a *Rokhlin tower* is a family  $\{F, TF, \dots, T^{n-1}F\}$  of pairwise disjoint sets, while “tending to the partition into points” means that we can approximate every set in  $\mathcal{B}$  by unions of levels of towers in the sequence). Baxter [20] showed that the maximal spectral type of such a  $T$  is realized by a characteristic function. Since the cyclic space generated by the characteristic function of the base contains characteristic functions of all levels of the tower, by the definition of rank one, the increasing sequence of cyclic spaces tends to the whole  $L^2$ -space, therefore rank one property implies simplicity of the spectrum for the Koopman representation. It was a question for some time whether rank-1 is just a characterization of simplicity of the spectrum, disproved by del Junco [88]. We refer the reader to [46] as a good source for basic properties of rank-1 transformations.

Similarly to the rank one property, one can define *finite rank* automorphisms (simply by requiring that an approximation is given by a sequence of a fixed number of towers) – see e.g. [152], or even, a more general property, namely the *local rank one* property can be defined, just by requiring that the approximating sequence of single towers fills up a fixed fraction of the space (see [44, 111]). Local rank one (so the more finite rank) property implies finite multiplicity [111]. Mentzen [146] showed that for each  $n \geq 1$  one can construct an automorphism with simple spectrum and having rank  $n$ ; in [138] there is an example of a simple spectrum automorphism which is not of local rank one. Ferenczi [45] introduced the notion of funny rank one (approximating towers are Rokhlin towers with “holes”). Funny rank one also implies simplicity of the spectrum. An example is given in [45] which is

even not loosely Bernoulli (see Sect. “[Inducing and Spectral Theory](#)”, we recall that local rank one property implies loose Bernoullicity [44]).

The notion of AT (see the glossary) has been introduced by Connes and Woods [25]. They proved that AT property implies zero entropy. They also proved that funny rank one automorphisms are AT. In [32] it is proved that the system induced by the classical Morse-Thue system is AT (it is an open question by S. Ferenczi whether this system has funny rank one property). A question by Dooley and Quas is whether AT implies funny rank one property. AT property implies “simplicity of the spectrum in  $L^1$ ” which we already considered in Introduction (a “generic” proof of this fact is due to J.-P. Thouvenot).

A persistent question was formulated in the 1980s whether rank one itself is a spectral property. In [49] the authors maintained that this is not the case, based on an unpublished preprint of the first named author of [49] in which there was a construction of a Gaussian–Kronecker automorphism (see Sect. “[Spectral Theory of Dynamical Systems of Probabilistic Origin](#)”) having rank-1 property. This latter construction turned out to be false. In fact de la Rue [181] proved that no Gaussian automorphism can be of local rank one. Therefore the question whether: *Rank one is a spectral property* remains one of the interesting open questions in that theory. Downarowicz and Kwiatkowski [33] proved that rank-1 is a spectral property in the class of systems generated by generalized Morse sequences.

One of the most beautiful theorems about rank-1 automorphisms is the following result of J. King [110] (for a different proof see [186]).

**Theorem 8 (WCT)** *If  $T$  is of rank one then for each element  $S$  of the centralizer  $C(T)$  of  $T$  there exists a sequence  $(n_k)$  such that  $U_T^{n_k} \rightarrow U_S$  strongly.*

A conjecture of J. King is that in fact for rank-1 automorphisms each indecomposable Markov operator  $J = J_\rho$  ( $\rho \in J_2^e(T)$ ) is a weak limit of powers of  $U_T$  (see [112], also [186]). To which extent the WCT remains true for actions of other groups is not clear. In [214] the WCT is proved in case of rank one flows, however the main argument seems to be based on the fact that a rank one flow has a non-zero time automorphism  $T_{t_0}$  which is of rank one, which is not true. After the proof of the WCT by Ryzhikov in [186] there is a remark that the rank one flow version of the theorem can be proved by a word for word repetition of the arguments. He also proves that if the flow  $(T_t)_{t \in \mathbb{R}}$  is mixing, then  $T_1$  does not have finite rank. On the other hand, for  $\mathbb{A} = \mathbb{Z}^2$ , Downarowicz and Kwiatkowski [34] gave recently a counterexample to the WCT.



Even though it looks as if rank one construction is not complicated, mixing in this class is possible; historically the first mixing constructions were given by D. Ornstein [151] in 1970, using probability type arguments for a choice of spacers. Once mixing was shown, the question arose whether absolutely continuous spectrum is also possible, as this would give automatically the positive answer to the Banach problem. However Bourgain [21], relating spectral measures of rank one automorphisms with some classical constructions of Riesz product measures, proved that a certain subclass of Ornstein's class consists of automorphisms with singular spectrum (see also [5] and [6]). Since in Ornstein's class spacers are chosen in a certain "non-constructive" way, quite a lot of attention was devoted to the rank one automorphism defined by cutting a tower at the  $n$ th step into  $r_n = n$  subcolumns of equal "width" and placing  $i$  spacers over the  $i$ th subcolumn. The mixing property conjectured by M. Smorodinsky, was proved by Adams [7] (in fact Adams proved a general result on mixing of a class of staircase transformations). Spectral properties of rank-1 transformations are also studied in [114], where the authors proved that whenever  $\sum_{n=1}^{\infty} r_n^{-2} = +\infty$  ( $r_n$  stands for the number of subcolumns at the  $n$ th step of the construction of a rank-1 automorphism) then the spectrum is automatically singular. H. Abdalaoui [5] gives a criterion for singularity of the spectrum of a rank one transformation; his proof uses a central limit theorem. It seems that still the question whether rank one implies singularity of the spectrum remains the most important question of this theory.

We have already seen in Sect. "Spectral Theory of Weighted Operators" that for a special class of rank one systems, namely those with discrete spectra ([87]), we have a nice theory for weighted operators. It would be extremely interesting to find a rank one automorphism with continuous spectrum for which a substitute of Helson's analysis exists.

B. Fayad [39] constructs a rank one differentiable flow, as a special flow over a two-dimensional rotation. In [40] he gives new constructions of smooth flows with singular spectra which are mixing (with a new criterion for a Rajchman measure to be singular). In [35] a certain smooth change of time for an irrational flows on the 3-torus is given, so that the corresponding flow is partially mixing and has the local rank one property.

### Spectral Theory of Dynamical Systems of Probabilistic Origin

Let us just recall that when  $(Y_n)_{n=-\infty}^{\infty}$  is a stationary process then its distribution  $\mu$  on  $\mathbb{R}^{\mathbb{Z}}$  is invariant un-

der the shift  $S$  on  $\mathbb{R}^{\mathbb{Z}}$ :  $S((x_n)_{n \in \mathbb{Z}}) = (y_n)_{n \in \mathbb{Z}}$ , where  $y_n = x_{n+1}$ ,  $n \in \mathbb{Z}$ . In this way we obtain an automorphism  $S$  defined on  $(\mathbb{R}^{\mathbb{Z}}, \mathcal{B}(\mathbb{R}^{\mathbb{Z}}), \mu)$ . For each automorphism  $T$  we can find  $f: X \rightarrow \mathbb{R}$  measurable such that the smallest  $\sigma$ -algebra making the stationary process  $(f \circ T^n)_{n \in \mathbb{Z}}$  measurable is equal to  $\mathcal{B}$ , therefore, for the purpose of this article, by a system of probabilistic origin we will mean  $(S, \mu)$  obtained from a stationary infinitely divisible process (see e.g. [142, 192]). In particular, the theory of Gaussian dynamical systems is indeed a classical part of ergodic theory (e.g. [149, 150, 211, 212]). If  $(X_n)_{n \in \mathbb{Z}}$  is a stationary real centered Gaussian process and  $\sigma$  denotes the *spectral measure of the process*, i.e.  $\widehat{\sigma}(n) = E(X_n \cdot X_0)$ ,  $n \in \mathbb{Z}$ , then by  $S = S_{\sigma}$  we denote the corresponding Gaussian system on the shift space (recall also that for each symmetric measure  $\sigma$  on  $\mathbb{T}$  there is exactly one stationary real centered Gaussian process whose spectral measure is  $\sigma$ ). Notice that if  $\sigma$  has an atom, then in the cyclic space generated by  $X_0$  there exists an eigenfunction  $Y$  for  $S_{\sigma}$  – if now  $S_{\sigma}$  were ergodic,  $|Y|$  would be a constant function which is not possible by the nature of elements in  $\mathbb{Z}(X_0)$ . In what follows we assume that  $\sigma$  is continuous.

It follows that  $U_{S_{\sigma}}$  restricted to  $\mathbb{Z}(X_0)$  is spectrally the same as  $V = V^{\sigma}$  acting on  $L^2(\mathbb{T}, \sigma)$ , and we obtain that  $(U_{S_{\sigma}}, L^2(\mathbb{R}^{\mathbb{Z}}, \mu_{\sigma}))$  can be represented as the symmetric Fock space built over  $H = L^2(\mathbb{T}, \sigma)$  and  $U_{S_{\sigma}} = F(V)$  – see the glossary ( $H^{\odot n}$  is called the  $n$ -th chaos). In other words the spectral theory of Gaussian dynamical systems is reduced to the spectral theory of special tensor products unitary operators. Classical results (see [26]) which can be obtained from this point of view are for example the following:

- (A) ergodicity implies weak mixing,
- (B) the multiplicity function is either 1 or is unbounded,
- (C) the maximal spectral type of  $U_{S_{\sigma}}$  is equal to  $\exp(\sigma)$ , hence Gaussian systems enjoy the Kolmogorov group property.

However we can also look at a Gaussian system in a different way, simply by noticing that the variables  $e^{2\pi i f}$  ( $f$  is a real variable), where  $f \in \mathbb{Z}(X_0)$  generate  $L^2(\mathbb{R}^{\mathbb{Z}}, \mu_{\sigma})$ . Now calculating the spectral measure of  $e^{2\pi i f}$  is not difficult and we obtain easily (C). Moreover, integrals of type  $\int e^{2\pi i f_0} e^{2\pi i f_1 \circ T^n} e^{2\pi i f_2 \circ T^{n+m}} d\mu_{\sigma}$  can also be calculated, whence in particular we easily obtain Leonov's theorem on the multiple mixing property of Gaussian systems [141].

One of the most beautiful parts of the theory of Gaussian systems concerns ergodic properties of  $S_{\sigma}$  when  $\sigma$  is concentrated on a thin Borel set. Recall that a closed sub-

set  $K \subset \mathbb{T}$  is said to be a *Kronecker set* if each  $f \in C(K)$  is a uniform limit of characters (restricted to  $K$ ). Each Kronecker set has no rational relations. Gaussian–Kronecker automorphisms are, by definition, those Gaussian systems for which the measure  $\sigma$  (always assumed to be continuous) is concentrated on  $K \cup \overline{K}$ ,  $K$  a Kronecker set. The following theorem has been proved in [51] (see also [26]).

**Theorem 9 (Foiş–Stratila Theorem)** *If  $T$  is an ergodic automorphism and  $f$  is a real-valued element of  $L^2_0$  such that the spectral measure  $\sigma_f$  is concentrated on  $K \cup \overline{K}$ , where  $K$  is a Kronecker set, then the process  $(f \circ T^n)_{n \in \mathbb{Z}}$  is Gaussian.*

This theorem is indeed striking as it gives examples of weakly mixing automorphisms which are spectrally determined (like rotations). A relative version of the Foiş–Stratila Theorem has been proved in [129].

The Foiş–Stratila Theorem implies that whenever a spectral measure  $\sigma$  is Kronecker, it has no realization of the form  $\sigma_f$  with  $f$  bounded. We will see however in Sect. “Future Directions” that for some automorphisms  $T$  (having the SCS property) the maximal spectral type  $\sigma_T$  has the property that  $\sigma_T$  has a simple spectrum.

Gaussian–Kronecker automorphisms are examples of automorphisms with simple spectra. In fact, whenever  $\sigma$  is concentrated on a set without rational relations, then  $\sigma_\sigma$  has a simple spectrum (see [26]). Examples of mixing automorphisms with simple spectra are known [149], however it is still unknown (Thouvenot’s question) whether the Foiş–Stratila property may hold in the mixing class. F. Parreau [154] using independent Helson sets gave an example of mildly mixing Gaussian system with the Foiş–Stratila property.

In [165] there is a remark that the set of finite essential values of the multiplicity function of  $U_{S_\sigma}$  forms a (multiplicative) subsemigroup of  $\mathbb{N}$ . However, it seems that there is no “written” proof of this fact.

Joining theory of a class of Gaussian system, called GAG, is developed in [136]. A Gaussian automorphism  $S_\sigma$  with the Gaussian space  $H \subset L^2_0(\mathbb{R}^{\mathbb{Z}}, \mu_\sigma)$  is called a GAG if for each ergodic self-joining  $\rho \in J^e_2(S_\sigma)$  and arbitrary  $f, g \in H$  the variable

$$(\mathbb{R}^{\mathbb{Z}} \times \mathbb{R}^{\mathbb{Z}}, \rho) \ni (x, y) \mapsto f(x) + g(y)$$

is Gaussian. For GAG systems one can describe the centralizer and factors, they turn out to be objects close to the probability structure of the system. One of the crucial observations in [136] was that all Gaussian systems with simple spectrum are GAG.

It is conjectured (J.P. Thouvenot) that in the class of zero entropy Gaussian systems the PID property holds true.

For the spectral theory of classical factors of a Gaussian system see [137]; also spectrally they share basic spectral properties of Gaussian systems. Recall that historically one of the classical factors namely the  $\sigma$ -algebra of sets invariant for the map

$$(\dots, x_{-1}, x_0, x_1, \dots) \mapsto (\dots, -x_{-1}, -x_0, -x_1, \dots)$$

was the first example with zero entropy and countable Lebesgue spectrum (indeed, we need a singular measure  $\sigma$  such that  $\sigma * \sigma$  is equivalent to Lebesgue measure [150]). For factors obtained as functions of a stationary process see [83].

T. de la Rue [181] proved that Gaussian systems are never of local rank-1, however his argument does not apply to classical factors. We conjecture that Gaussian systems are disjoint from rank-1 automorphisms (or even from local rank-1 systems).

We now turn the attention to Poissonian systems (see [26] for more details). Assume that  $(X, \mathcal{B}, \mu)$  is a standard Borel space, where  $\mu$  is infinite,  $\sigma$ -finite. The new configuration space  $\tilde{X}$  is taken as the set of all countable subsets  $\{x_i : i \geq 1\}$  of  $X$ . Once a set  $A \in \mathcal{B}$ , of finite measure is given one can define a map  $N_A : \tilde{X} \rightarrow \mathbb{N}(\cup \{\infty\})$  just counting the number of elements belonging to  $A$ . The measure-theoretic structure  $(\tilde{X}, \tilde{\mathcal{B}}, \tilde{\mu})$  is given so that the maps  $N_A$  become random variables with Poisson distribution of parameter  $\mu(A)$  and such that whenever  $A_1, \dots, A_k \subset X$  are of finite measure and are pairwise disjoint then the variables  $N_{A_1}, \dots, N_{A_k}$  are independent.

Assume now that  $T$  is an automorphism of  $(X, \mathcal{B}, \mu)$ . It induces a natural automorphism on the space  $(\tilde{X}, \tilde{\mathcal{B}}, \tilde{\mu})$  defined by  $\tilde{T}(\{x_i : i \geq 1\}) = \{Tx_i : i \geq 1\}$ . The automorphism  $\tilde{T}$  is called the *Poisson suspension* of  $T$  (see [26]). Such a suspension is ergodic if and only if no set of positive and finite  $\mu$ -measure is  $T$ -invariant. Moreover ergodicity of  $\tilde{T}$  implies weak mixing. In fact the spectral structure of  $U_{\tilde{T}}$  is very similar to the Gaussian one: namely the first chaos equals  $L^2(X, \mathcal{B}, \mu)$  (we emphasize that this is about the whole  $L^2$  and not only  $L^2_0$ ) on which  $U_{\tilde{T}}$  acts as  $U_T$  and the  $L^2(\tilde{X}, \tilde{\mu})$  together with the action of  $U_{\tilde{T}}$  has the structure of the symmetric Fock space  $F(L^2(X, \mathcal{B}, \mu))$  (see the glossary).

We refer to [22,86,168,169] for ergodic properties of systems given by symmetric  $\alpha$ -stable stationary processes, or more generally infinitely divisible processes. Again, they share spectral properties similar to the Gaussian case: er-

godicity implies weak mixing, while mixing implies mixing of all orders.

In [171], E. Roy clarifies the dynamical “status” of such systems. He uses Poisson suspension automorphisms and the Maruyama representation of an infinitely divisible process mixed with basic properties of automorphisms preserving infinite measure (see [1]) to prove that as a dynamical system, a stationary infinitely divisible process (without the Gaussian part), is a factor of the Poisson suspension over the Lévy measure of this process. In [170] a theory of ID-joinings is developed (which should be viewed as an analog of the GAG theory in the Gaussian class). Parreau and Roy [155] give an example of a Poisson suspension with a minimal possible set of ergodic self-joinings.

Many natural problems still remain open here, for example (assuming always zero entropy of the dynamical system under consideration): Are Poisson suspensions disjoint from Gaussian systems? What is the spectral structure for dynamical systems generated by symmetric  $\alpha$ -stable process? Are such systems disjoint whenever  $\alpha_1 \neq \alpha_2$ ? Are Poissonian systems disjoint from local rank one automorphisms (cf. [181])? In [91] it is proved that Gaussian systems are disjoint from so called simple systems (see ► [Joinings in Ergodic Theory](#) and [93,208]); we will come back to an extension of this result in Sect. “[Future Directions](#)”. It seems that flows of probabilistic origin satisfy the Kolmogorov group property for the spectrum. One can therefore ask how different are systems satisfying the Kolmogorov group property from systems for which the convolutions of the maximal spectral type are pairwise disjoint (see also Sect. “[Future Directions](#)” and the SCS property).

We also mention here another problem which will be taken up in Sect. “[Special Flows and Flows on Surfaces, Interval Exchange Transformations](#)” – *Is it true that flows of probabilistic origin are disjoint from smooth flows on surfaces?* Recently A. Katok and A. Windsor announced that it is possible to construct a Kronecker measure so that the corresponding Gaussian system ( $\mathbb{Z}$ -action (!)) has a smooth representation on the torus.

Yet one more (joining) property seems to be characteristic in the class of systems of probabilistic origin, namely they satisfy so called ELF property (see [30] and ► [Joinings in Ergodic Theory](#)). Vershik asked whether the ELF property is spectral – however the example of a cocycle from [205] together with Theorem 7 (i) yields a certain Rokhlin extension of a rotation which is ELF and has countable Lebesgue spectrum in the orthocomplement of the eigenfunctions (see [206]); on the other hand any affine extension of that rotation is spectrally the same, while it cannot have the ELF property.

Prikhodko and Thouvenot (private communication) have constructed weakly mixing and non-mixing rank one automorphisms which enjoy the ELF property.

## Inducing and Spectral Theory

Assume that  $T$  is an ergodic automorphism of a standard probability Borel space  $(X, \mathcal{B}, \mu)$ . Can “all” dynamics be obtained by inducing (see the glossary) from one fixed automorphism was a natural question from the very beginning of ergodic theory. Because of Abramov’s formula for entropy  $h(T_A) = h(T)/\mu(A)$  it is clear that positive entropy transformations cannot be obtained from inducing on a zero entropy automorphism. However here we are interested in spectral questions and thus we ask how many spectral types we obtain when we induce. It is proved in [59] that the family of  $A \in \mathcal{B}$  for which  $T_A$  is mixing is dense for the (pseudo) metric  $d(A_1, A_2) = \mu(A_1 \triangle A_2)$ . De la Rue [182] proves the following result: *For each ergodic transformation  $T$  of a standard probability space  $(X, \mathcal{B}, \mu)$  the set of  $A \in \mathcal{B}$  for which the maximal spectral type of  $U_{T_A}$  is Lebesgue is dense in  $\mathcal{B}$ .* The multiplicity function is not determined in that paper. Recall (without giving a formal definition, see [152]) that a zero entropy automorphism is *loosely Bernoulli* (LB for short) if and only if it can be induced from an irrational rotation (see also [43,99]). The LB theory shows that not all dynamical systems can be obtained by inducing from an ergodic rotation. However an open question remained whether LB systems exhaust spectrally all Koopman representations. In a deep paper [180], de la Rue studies LB property in the class of Gaussian–Kronecker automorphisms, in particular he constructs  $S$  which is not LB. Suppose now that  $T$  is LB and for some  $A \in \mathcal{B}$ ,  $U_{T_A}$  is isomorphic to  $U_S$ . Then by the Foiaş–Stratila Theorem,  $T_A$  is isomorphic to  $S$ , and hence  $T_A$  is not LB. However an induced automorphism from an LB automorphism is LB, a contradiction.

## Special Flows and Flows on Surfaces, Interval Exchange Transformations

We now turn our attention to flows. The cases of the geodesic flow, horocycle flows on homogenous spaces of  $SL(2, \mathbb{R})$  and nilflows are classical (we refer the reader to [105] with a nice description of the first two cases, while for nilflows we refer to [157]: these classes of flows on homogenous spaces have countable Lebesgue spectrum, in the third case – in the orthocomplement of the eigenspace). On the other hand the classical cyclic approximation theory of Katok and Stepin [104] (see [26]) leads to examples of smooth flows on the torus with simple continuous singular spectra.

Given an ergodic automorphism  $T$  on  $(X, \mathcal{B}, \mu)$  and a positive integrable function  $f: X \rightarrow \mathbb{R}^+$  consider the corresponding special flow  $T^f$  (see the glossary). Obviously, such a flow is ergodic. Special flows were introduced to ergodic theory by von Neumann in his fundamental work [148] in 1932. Also in that work he explains how to compute eigenvalues for special flows, namely:  $r \in \mathbb{R}$  is an eigenvalue of  $T^f$  if and only if the following functional equation

$$e^{2\pi i r f(x)} = \frac{\xi(x)}{\xi(Tx)}$$

has a measurable solution  $\xi: X \rightarrow \mathbb{T}$ . We recall also that the classical Ambrose-Kakutani theorem asserts that practically each ergodic flow has a special representation ([26], see also Rudolph's theorem on special representation therein).

A classical situation when we obtain "natural" special representations is while considering smooth flows on surfaces (we refer the reader to Hasselblatt's and Katok's monograph [73]). They have transversals on which the Poincaré map is piecewise isometric, and this entails a study of interval exchange transformations (IET), see [26, 108, 163]. Formally, to define IET of  $m$  intervals we need a permutation  $\pi$  of  $\{1, \dots, m\}$  and a probability vector  $\lambda = (\lambda_1, \dots, \lambda_m)$  (with positive entries). Then we define  $T = T_{\lambda, \pi}$  of  $[0, 1)$  by putting

$$T_{\lambda, \pi}(x) = x + \beta_i^\pi - \beta_i \text{ for } x \in [\beta_i, \beta_{i+1}),$$

where  $\beta_i = \sum_{j < i} \lambda_j$ ,  $\beta_i^\pi = \sum_{\pi(j) < \pi(i)} \beta_j$ . Obviously, each IET preserves Lebesgue measure. One of possible approaches to study ergodic properties of IET is an "a.e." approach "seen" in the definition of  $T_{\lambda, \pi}$ . It is based on the fundamental fact that the induced transformation on a subinterval of  $[0, 1)$  is also IET (see [26]). This leads to a very delicate and deep mathematics based on Rauzy induction, which is a way of inducing on special intervals, considering only irreducible permutations whose set is partitioned into orbits of some maps (any such an orbit is called a *Rauzy class*). If now  $\mathcal{R}$  is a Rauzy class of permutations and  $\lambda$  lies in the standard simplex  $\Delta_{m-1}$  then the Rauzy induction together with a natural renormalization leads to a map  $\mathcal{P}: \mathcal{R} \times \Delta_{m-1} \rightarrow \mathcal{R} \times \Delta_{m-1}$ . For a better understanding of the dynamics of the Rauzy map Veech [209] introduced the space of *zippered rectangles*. A zippered rectangle associated to the Rauzy class  $\mathcal{R}$  is a quadruple  $(\lambda, h, a, \pi)$ , where  $\lambda \in \mathbb{R}_+^m$ ,  $h \in \mathbb{R}_+^m$ ,  $a \in \mathbb{R}_+^m$ ,  $\pi \in \mathcal{R}$  and the vectors  $h$  and  $a$  satisfy some equations and inequalities. Every zippered rectangle  $(\lambda, h, a, \pi)$  determines a Riemann structure on a compact connected sur-

face. Denote by  $\Omega(\mathcal{R})$  the space of all zippered rectangles, corresponding to a given Rauzy class  $\mathcal{R}$  and satisfying the condition  $\langle \lambda, h \rangle = 1$ . In [209], Veech defined a flow  $(P^t)_{t \in \mathbb{R}}$  on the space  $\Omega(\mathcal{R})$  putting

$$P^t(\lambda, h, a, \pi) = (e^t \lambda, e^{-t} h, e^{-t} a, \pi)$$

and extended the Rauzy map. On so called *Veech moduli space* of zippered rectangles, the flow  $(P^t)$  becomes the *Teichmüller flow* and it preserves a natural Lebesgue measure class; by passing to a transversal and projecting the measure on the space of IETs  $\mathcal{R} \times \Delta_{m-1}$  Veech has proved the following fundamental theorem ([209], see also [143]) which is a generalization of the fact that Gauss measure  $1/(\ln 2)1/(1+x)dx$  is invariant for the Gauss map which sends  $t \in (0, 1)$  into the fractional part of its inverse.

**Theorem 10 (Veech, Masur, 1982)** *Assume that  $\mathcal{R}$  is a Rauzy class. There exists a  $\sigma$ -finite measure  $\mu_{\mathcal{R}}$  on  $\mathcal{R} \times \Delta_{m-1}$  which is  $\mathcal{P}$ -invariant, equivalent to "Lebesgue" measure, conservative and ergodic.*

In [209] it is proved that a.e. (in the above sense) IET is then of rank one (and hence is ergodic and has a simple spectrum). He also formulated as an open problem whether we can achieve the weak mixing property a.e. This has been recently answered in positive by A. Avila and G. Forni [19] (for  $\pi$  which is not a rotation).

Katok [100] proved that IET have no mixing factors (in fact his proof shows more: the IET class is disjoint with the class of mixing transformations). By their nature, IET transformations are of finite rank (see [26]) so they are of finite multiplicity. They need not be of simple spectrum (see remarks in [105] pp. 88–90). It remains an open question whether an IET can have a non-singular spectrum. Joining properties in the class of exchange of 3 and more intervals are studied in [47, 48]. An important question of Veech [208] whether a.e. IET is simple is still open.

When we consider a smooth flow on a surface preserving a smooth measure, whose only singularity (we assume that we have only finitely many singularities) are simple (non-degenerated) saddles then such a flow has a special representation over an interval exchange automorphism under a smooth function which has finitely many logarithmic singularities (see [73]). In the article by Arnold [18] the quasi-periodic Hamiltonian case is considered:  $H: \mathbb{R}^2 \rightarrow \mathbb{R}$  satisfies  $H(x+m, y+n) = H(x, y) + n\alpha_1 + m\alpha_2$ ,  $\alpha_1/\alpha_2 \notin \mathbb{Q}$ , and we then consider the following system of differential equations on  $\mathbb{T}^2$

$$\frac{dx}{dt} = \frac{\partial H}{\partial y}, \quad \frac{dy}{dt} = -\frac{\partial H}{\partial x}. \quad (2)$$



As Arnold shows the dynamical system arising from the system (2) has one ergodic component which has a special representation over the irrational rotation by  $\alpha := \alpha_1/\alpha_2$  under a smooth function with finitely many logarithmic singularities (all other ergodic components are periodic orbits and separatrices). By changing a speed as it is done in [57] so that critical points of the vector field in (2) become singular points, Arnold's special representation is transformed to a special flow over the same irrational rotation however under a piecewise smooth function. If the sum of jumps is not zero then in fact we come back to von Neumann's class of special flows considered in [148]. Similar classes of special flows (when the roof function is of bounded variation) are obtained from ergodic components of flows associated to billiards in convex polygons with rational angles [106]. Kochergin [115] showed that special flows over irrational rotations and under bounded variation functions are never mixing. This has been recently strengthened in [54] to the following: *If  $T$  is an irrational rotation and  $f$  is of bounded variation then the special flow  $T^f$  is spectrally disjoint from all mixing flows.* In particular all such flows have singular spectra. Moreover, in [54] it is proved that whenever the Fourier transform of the roof function  $f$  is of order  $O(1/n)$  then  $T^f$  is disjoint from all mixing flows (see also [55]). In fact in the papers [54,55,56,57] the authors discuss the problem of disjointness of those special flows with all ELF-flows conjecturing that no flow of probabilistic origin has a smooth realization on a surface. In [140] the analytic case is considered leading to a "generic" result on disjointness with the ELF class generalizing the classical Shklover's result on the weak mixing property [197].

Kochergin [117] proved the absence of mixing for flows where the roof function has finitely many singularities, whenever the sum of "left logarithmic speeds" and the sum of "right logarithmic speeds" are equal – this is called a *symmetric logarithmic case*, however some Diophantine restriction is put on  $\alpha$ .

In [128], where also the absence of mixing is considered for the symmetric logarithmic case, it was conjectured (and proved for arbitrary rotation) that a necessary condition for mixing of a special flow  $T^f$  (with arbitrary  $T$  and  $f$ ) is the condition that the sequence of distributions  $((f_0^{(n)})_*)_n$  tends to  $\delta_\infty$  in the space of probability measures on  $\mathbb{R}$ . K. Schmidt [194] proved it using the theory of cocycles and extending a result from [3] on tightness of cocycles.

A. Katok [100] proved the absence of mixing for special flows over IET when the roof function is of bounded variation (see also [187]). Katok's theorem was strength-

ened in [56] to the disjointness theorem with the class of mixing flows.

On the other hand there is a lot of (difficult) results pointing out classes of special flows over irrational rotations which are mixing, especially (but not only) in the class of non-symmetric logarithmic singularities: [36,38] (B. Fayad was able to give a speed of convergence to zero for Fourier coefficients), [109,119,120]. Recently mixing property has been proved in a non-symmetric case in [203] when the base transformation is a special class of IETs.

The eigenvalue problem (mainly how many frequencies can have the group of eigenvalues) for special flows over irrational rotations is studied in [41,42,71].

A. Avila and G. Forni [19] proved that a.e. translation flow on a surface (of genus at least two) is weakly mixing (which is a drastic difference with the linear flow case of the torus, where the spectrum is always discrete).

The problem of whether mixing flows indicated in this chapter are mixing of all orders is open (it is also unknown whether they have singular spectra). One of several possible approaches (proposed by B. Fayad and J.-P. Thouvenot) toward positive solution of this problem would be to show that such flows enjoy so called Ratner's property (R-property). This property may be viewed as a particular way of divergence of orbits of close points; it was shown to hold for horocycle flows by M. Ratner [162]. We refer the reader to [162] and the survey article [201] for the formal definitions and basic consequences of R-property. In particular, R-property implies "rigidity" of joinings and it also implies the PID property; hence mixing and R-property imply mixing of all orders. In [57,58] a version of R-property is shown for the class of von Neumann special flows (however  $\alpha$  is assumed to have bounded partial quotients). This allowed one to prove there that such flows are even mildly mixing (mixing is excluded by a Kochergin's result). We conjecture that an R-property may also hold for special flows over multidimensional rotations with roof functions given by nil-cocycles which we mentioned in Sect. "Spectral Theory of Weighted Operators".

If indeed the R-property is ubiquitous in the class of smooth flows on surfaces it may also be useful to show that smooth flows on surfaces are disjoint with flows of probabilistic origin – see [91,92,135,190,202].

B. Fayad [40] gives a criterion that implies singularity of the maximal spectral type for a dynamical system on a Riemannian manifold. As an application he gives a class of smooth mixing flows (with singular spectra) on  $\mathbb{T}^3$  obtained from linear flows by a time change (again this is a drastic difference with dimension two, where a smooth time change of a linear flow leads to non-mixing flows [26]).



The spectral multiplicity problem for special flows (with sufficiently regular roof functions) over irrational rotations seems to be completely untouched (except for the case of a sufficiently smooth  $f$  – the spectrum of  $T^f$  is then simple [26]). It would be nice to have examples of such flows with finite bigger than one multiplicity. In particular, is it true that the von Neumann class of special flows have finite multiplicity? This was partially solved by A. Katok (private communication) on certain subspaces in  $L^2$ , but not on the whole  $L^2$ -space.

**Problem.** Given  $Tx = x + \alpha$  (with  $\alpha$  irrational) can we find  $f: [0, 1) \rightarrow \mathbb{R}^+$  sufficiently regular (e.g. with finitely many “controllable” singularities) such that  $T^f$  has a Lebesgue spectrum?

Of course the above is related to the question whether at all one can find a smooth flow on a surface with a Lebesgue spectrum (for  $\mathbb{Z}$ -actions we can even see positive entropy diffeomorphisms on the torus).

We mention at the end that if we drop here (and in other problems) the assumption of regularity of  $f$  then the answers will be always positive because of the LB theory; in particular there is a section of any horocycle flow (it has the LB property [161]) such that in the corresponding special representation  $T^f$  the map  $T$  is an irrational rotation. Using a Kochergin’s result [118] on cohomology (see also [98, 176]) the  $L^1$ -function  $f$  is cohomologous to a positive function  $g$  which is even continuous, thus  $T^f$  is isomorphic to  $T^g$ .

## Future Directions

We have already seen several cases where spectral properties interact with measure-theoretic properties of a system. Let us mention a few more cases which require further research and deeper understanding.

We recall that the weak mixing property can be understood as a property complementary to discrete spectrum (more precisely to the distality [62]), or similarly mild mixing property is complementary to rigidity. This can be phrased quite precisely by saying that  $T$  is not weakly (mildly) mixing if and only if it has a non-trivial factor with discrete spectrum (it has a non-trivial rigid factor). It has been a question for quite a long time if in a sense mixing can be “built” on the same principle. In other words we seek a certain “highly” non-mixing factor. It was quite surprising when in 2005 F. Parreau (private communication) gave the positive answer to this problem.

**Theorem 11 (Parreau)** Assume that  $T$  is an ergodic automorphism of a standard probability space  $(X, \mathcal{B}, \mu)$ . Assume moreover that  $T$  is not mixing. Then there exists

a non-trivial factor (see below) of  $T$  which is disjoint with all mixing automorphisms.

In fact, Parreau proved that each factor of  $T$  given by  $\mathcal{B}_\infty(\rho)$  (this  $\sigma$ -algebra is described in [136]), where  $U_T^{n_k} \rightarrow J_\rho$ , is disjoint from all mixing transformations. This proof leads to some other results of the same type, for example: Assume that  $T$  is an ergodic automorphism of a standard probability space. Assume that there exists a non-trivial automorphism  $S$  with a singular spectrum which is not disjoint with  $T$ . Then  $T$  has a non-trivial factor which is disjoint with any automorphism with a Lebesgue spectrum.

The problem of spectral multiplicity of Cartesian products for “typical” transformation studied by Katok [98] and then its solution in [13] which we already considered in Sect. “The Multiplicity Function” lead to a study of those  $T$  for which

$$(CS) \quad \sigma^{(m)} \perp \sigma^{(n)} \text{ whenever } m \neq n,$$

where  $\sigma = \sigma_T$  just stands for the reduced maximal spectral type of  $U_T$  (which is constantly assumed to be a continuous measure), see also Stepin’s article [199].

The usefulness of the above property (CS) in ergodic theory was already shown in [90], where a spectral counterexample machinery was presented using the following observation: If  $\mathcal{A}$  is a  $T^{\times\infty}$ -invariant sub- $\sigma$ -algebra such that the maximal spectral type on  $L^2(\mathcal{A})$  is absolutely continuous with respect to  $\sigma_T$  then  $\mathcal{A}$  is contained in one of the coordinate sub- $\sigma$ -algebras  $\mathcal{B}$ . Based on that in [90] it is shown how to construct two weakly isomorphic action which are not isomorphic or how to construct two non-disjoint automorphisms which have no common non-trivial factors (such constructions were previously known for so called minimal self-joining automorphisms [174]). See also [200] for extensions of those results to  $\mathbb{Z}^d$ -actions.

Prikhodko and Ryzhikov [159] proved that the classical Chacon transformation enjoys the (CS) property. The SCS property defined in the glossary is stronger than the (CS) condition above; the SCS property implies that the corresponding Gaussian system  $S_{\sigma_T}$  has a simple spectrum. Ageev [10] shows that Chacon’s transformation satisfies the SCS property; moreover in [13] he shows that the SCS property is satisfied generically and he gives a construction of a rank one mixing SCS-system (see also [191]). In [134] it is proved that some special flows considered in Sect. “Special Flows and Flows on Surfaces, Interval Exchange Transformations” (including the von Neumann class, however with  $\alpha$  having unbounded partial quotients) have the SCS property. Since the corresponding Gaussian systems have simple spectra, it would be interesting

to decide whether  $\sigma_T$  (for an SCS-automorphism) can be concentrated on a set without rational relations. It is quite plausible that the SCS property is commonly seen for smooth flows on surfaces.

Katok and Thouvenot (private communication) considered systems called *infinitely divisible*. These are systems  $T$  on  $(X, \mathcal{B}, \mu)$  which have a family of factors  $\mathcal{B}_\omega$  indexed by  $\omega \in \bigcup_{n=0}^{\infty} \{0, 1\}^n$  ( $\mathcal{B}_\varepsilon = \mathcal{B}$ ) such that  $\mathcal{B}_{\omega 0} \perp \mathcal{B}_{\omega 1}$ ,  $\mathcal{B}_{\omega 0} \vee \mathcal{B}_{\omega 1} = \mathcal{B}_\omega$  and for each  $\eta \in \{0, 1\}^{\mathbb{N}}$ ,  $\bigcap_{n \in \mathbb{N}} \mathcal{B}_{\eta[0, n]} = \{\emptyset, X\}$ . They showed (unpublished) that there are discrete spectrum transformations which are ID, and that there are rank one transformations with continuous spectra which are also ID (clearly Gaussian systems are ID). It was until recently that a relationship between ID automorphisms and systems coming from stationary ID processes was unclear. In [135] it is proved that dynamical systems coming from stationary ID processes are factors of ID automorphisms; moreover, ID automorphisms are disjoint with all systems having the SCS property. It would be nice to decide whether Koopman representations associated to ID automorphisms satisfy the Kolmogorov group property.

## Acknowledgments

Research supported by the EU Program Transfer of Knowledge “Operator Theory Methods for Differential Equations” TODEQ and the Polish Ministry of Science and High Education.

## Bibliography

- Aaronson J (1997) An introduction to infinite ergodic theory, mathematical surveys and monographs. Am Math Soc 50:284
- Aaronson J, Nadkarni MG (1987)  $L^\infty$  eigenvalues and  $L^2$  spectra of nonsingular transformations. Proc London Math Soc 55(3):538–570
- Aaronson J, Weiss B (2000) Remarks on the tightness of cocycles. Coll Math 84/85:363–376
- Abdalaoui H (2000) On the spectrum of the powers of Ornstein transformations Ergodic theory and harmonic analysis (Mumbai, 1999). Sankhya Ser. A 62:291–306
- Abdalaoui H (2007) A new class of rank 1 transformations with singular spectrum. Ergodic Theor Dynam Syst 27:1541–1555
- Abdalaoui H, Parreau F, Prikhodko A (2006) A new class of Ornstein transformations with singular spectrum. Ann Inst H Poincaré Probab Statist 42:671–681
- Adams T (1998) Smorodinsky's conjecture on rank one systems. Proc Am Math Soc 126:739–744
- Ageev ON (1988) Dynamical systems with a Lebesgue component of even multiplicity in the spectrum. Mat Sb (NS) 136(178):307–319, (Russian) 430 (1989) Translation in Math. USSR-Sb. 64:305–317
- Ageev ON (1999) On ergodic transformations with homogeneous spectrum. J Dynam Control Syst 5:149–152
- Ageev ON (2000) On the spectrum of cartesian powers of classical automorphisms. Mat Zametki 68:643–647, (2000) Translation Math. Notes 68:547–551, (Russian)
- Ageev ON (2001) On the multiplicity function of generic group extensions with continuous spectrum. Ergodic Theory Dynam Systems 21:321–338
- Ageev ON (2005) The homogeneous spectrum problem in ergodic theory. Invent Math 160:417–446
- Ageev ON (2007) Mixing with staircase multiplicity function. preprint
- Alexeyev VM (1958) Existence of bounded function of the maximal spectral type. Vestnik Mosc Univ 5:13–15 and (1982) Ergodic Theory Dynam Syst 2:259–261
- Anosov DV (2003) On Spectral multiplicities in ergodic theory. Institut im. V.A. Steklova, Moscow
- Anzai H (1951) Ergodic skew product transformations on the torus. Osaka J Math 3:83–99
- Arni N (2005) Spectral and mixing properties of actions of amenable groups. Electron Res Announc AMS 11:57–63
- Arnold VI (1991) Topological and ergodic properties of closed 1-forms with incommensurable periods. Funktsional Anal Prilozhen 25:1–12, (Russian)
- Avila A, Forni G (2007) Weak mixing for interval exchange transformations and translation flows. Ann Math 165:637–664
- Baxter JR (1971) A class of ergodic transformations having simple spectrum. Proc Am Math Soc 27:275–279
- Bourgain J (1993) On the spectral type of Ornstein's class of transformations. Isr J Math 84:53–63
- Cambanis S, Podgórski K, Weron A (1995) Chaotic behaviour of infinitely divisible processes. Studia Math 115:109–127
- Chacon RV (1970) Approximation and spectral multiplicity. In: Dold A, Eckmann B (eds) Contributions to ergodic theory and probability. Springer, Berlin, pp 18–27
- Choe GH Products of operators with singular continuous spectra. In: Operator theory: operator algebras and applications, Part 2. (Durham, 1988), pp 65–68, Proc Sympos Pure Math, 51, Part 2, Amer Math Soc, Providence, 1990
- Connes A, Woods E (1985) Approximately transitive flows and ITPFI factors. Ergod Theory Dynam Syst 5:203–236
- Cornfeld IP, Fomin SV, Sinai YG (1982) Ergodic theory. Springer, New York
- Danilenko AL (2006) Explicit solution of Rokhlin's problem on homogeneous spectrum and applications. Ergod Theory Dyn Syst 26:1467–1490
- Danilenko AL, Lemańczyk M (2005) A class of multipliers for  $\mathcal{W}^\perp$ . Isr J Math 148:137–168
- Danilenko AI, Park KK (2002) Generators and Bernouillian factors for amenable actions and cocycles on their orbits. Ergod Theory Dyn Syst 22:1715–1745
- Derriennic Y, Frączek K, Lemańczyk M, Parreau F (2007) Ergodic automorphisms whose weak closure of off-diagonal measures consists of ergodic self-joinings. Coll Math 110:81–115
- Dooley A, Golodets VY (2002) The spectrum of completely positive entropy actions of countable amenable groups. J Funct Anal 196:1–18
- Dooley A, Quas A (2005) Approximate transitivity for zero-entropy systems. Ergod Theory Dyn Syst 25:443–453

33. Downarowicz T, Kwiatkowski J (2000) Spectral isomorphism of Morse flows. *Fundam Math* 163:193–213
34. Downarowicz T, Kwiatkowski J (2002) Weak Cosure Theorem fails for  $\mathbb{Z}^d$ -actions. *Stud Math* 153:115–125
35. Fayad B (2001) Partially mixing and locally rank 1 smooth transformations and flows on the torus  $\mathbb{T}^d$ ,  $d \geq 3$ . *J London Math Soc* 64(2):637–654
36. Fayad B (2001) Polynomial decay of correlations for a class of smooth flows on the two torus. *Bull Soc Math France* 129:487–503
37. Fayad B (2002) Skew products over translations on  $T^d$ ,  $d \geq 2$ . *Proc Amer Math Soc* 130:103–109
38. Fayad B (2002) Analytic mixing reparametrizations of irrational flows. *Ergod Theory Dyn Syst* 22:437–468
39. Fayad B (2005) Rank one and mixing differentiable flows. *Invent Math* 160:305–340
40. Fayad B (2006) Smooth mixing flows with purely singular spectra. *Duke Math J* 132:371–391
41. Fayad B, Katok AB, Windsor A (2001) Mixed spectrum reparametrizations of linear flows on  $\mathbb{T}^2$ , dedicated to the memory of I. G. Petrovskii on the occasion of his 100th anniversary. *Mosc Math J* 1:521–537
42. Fayad B, Windsor A (2007) A dichotomy between discrete and continuous spectrum for a class of special flows over rotations. *J Mod Dyn* 1:107–122
43. Feldman J (1976) New  $K$ -automorphisms and a problem of Kakutani. *Isr J Math* 24:16–38
44. Ferenczi S (1984) Systèmes localement de rang un. *Ann Inst H Poincaré Probab Statist* 20:35–51
45. Ferenczi S (1985) Systèmes de rang un gauche. *Ann Inst H Poincaré Probab Statist* 21:177–186
46. Ferenczi S (1997) Systems of finite rank. *Colloq Math* 73:35–65
47. Ferenczi S, Holton C, Zamboni LQ (2005) Joinings of three-interval exchange transformations. *Ergod Theory Dyn Syst* 25:483–502
48. Ferenczi S, Holton C, Zamboni LQ (2004) Structure of three-interval exchange transformations III: ergodic and spectral properties. *J Anal Math* 93:103–138
49. Ferenczi S, Lemańczyk M (1991) Rank is not a spectral invariant. *Stud Math* 98:227–230
50. Filipowicz I (1997) Product  $\mathbb{Z}^d$ -actions on a Lebesgue space and their applications. *Stud Math* 122:289–298
51. Foiaş C, Stratila S (1968) Ensembles de Kronecker dans la théorie ergodique. *CR Acad Sci Paris, Ser A-B* 267:A166–A168
52. Frączek K (1997) On a function that realizes the maximal spectral type. *Stud Math* 124:1–7
53. Frączek K (2000) Circle extensions of  $\mathbb{Z}^d$ -rotations on the  $d$ -dimensional torus. *J London Math Soc* 61(2):139–162
54. Frączek K, Lemańczyk M (2004) A class of special flows over irrational rotations which is disjoint from mixing flows. *Ergod Theory Dyn Syst* 24:1083–1095
55. Frączek K, Lemańczyk M (2003) On symmetric logarithm and some old examples in smooth ergodic theory. *Fund Math* 180:241–255
56. Frączek K, Lemańczyk M (2005) On disjointness properties of some smooth flows. *Fund Math* 185:117–142
57. Frączek K, Lemańczyk M (2006) On mild mixing of special flows over irrational rotations under piecewise smooth functions. *Ergod Theory Dyn Syst* 26:719–738
58. Frączek K, Lemańczyk M, Lesigne E (2007) Mild mixing property for special flows under piecewise constant functions. *Disc Contin Dyn Syst* 19:691–710
59. Friedman NA, Ornstein DS (1973) Ergodic transformations induce mixing transformations. *Adv Math* 10:147–163
60. Furstenberg H (1967) Disjointness in ergodic theory, minimal sets, and a problem of Diophantine approximation. *Math Syst Theory* 1:1–49
61. Furstenberg H, Weiss B (1978) The finite multipliers of infinite ergodic transformations. *Lect Notes Math* 668:127–132
62. Furstenberg H (1981) Recurrence in ergodic theory and combinatorial number theory. Princeton University Press, Princeton
63. Glasner E (1994) On the class of multipliers for  $\mathcal{W}^\perp$ . *Ergod Theory Dyn Syst* 14:129–140
64. Glasner E (2003) Ergodic theory via joinings. *Math Surv Monogr*, AMS, Providence, RI 101:384
65. Glasner E, Weiss B (1989) Processes disjoint from weak mixing. *Trans Amer Math Soc* 316:689–703
66. Goodson GR (1999) A survey of recent results in the spectral theory of ergodic dynamical systems. *J Dyn Control Syst* 5:173–226
67. Goodson GR, Kwiatkowski J, Lemańczyk M, Liardet P (1992) On the multiplicity function of ergodic group extensions of rotations. *Stud Math* 102:157–174
68. Gromov AL (1991) Spectral classification of some types of unitary weighted shift operators. *Algebra Anal* 3:62–87, (Russian) (1992) Translation in *St. Petersburg Math J* 3:997–1021
69. Guenais M (1998) Une majoration de la multiplicité spectrale d'opérateurs associés à des cocycles réguliers. *Isr J Math* 105:263–284
70. Guenais M (1999) Morse cocycles and simple Lebesgue spectrum. *Ergod Theory Dyn Syst* 19:437–446
71. Guenais M, Parreau F (2005) Valeurs propres de transformations liées aux rotations irrationnelles et aux fonctions en escalier. (in press)
72. Hahn F, Parry W (1968) Some characteristic properties of dynamical systems with quasi-discrete spectra. *Math Syst Theory* 2:179–190
73. Hasselblatt B, Katok AB (2002) Principal structures. In: *Handbook of dynamical systems*, vol 1A. North-Holland, Amsterdam, pp 1–203
74. Helson H (1986) Cocycles on the circle. *J Oper Theory* 16:189–199
75. Helson H, Parry W (1978) Cocycles and spectra. *Arkiv Math* 16:195–206
76. Herman M (1979) Sur la conjugaison différentiable des difféomorphismes du cercle à des rotations. *Publ Math IHES* 49:5–234
77. Host B (1991) Mixing of all orders and pairwise independent joinings of systems with singular spectrum. *Isr J Math* 76:289–298
78. Host B, Méla F, Parreau F (1991) Non-singular transformations and spectral analysis of measures. *Bull Soc Math France* 119:33–90
79. Iwanik A (1991) The problem of  $L^p$ -simple spectrum for ergodic group automorphisms. *Bull Soc Math France* 119:91–96
80. Iwanik A (1992) Positive entropy implies infinite  $L^p$ -multiplicity for  $p > 1$ . *Ergod Theory Rel Top* (Güstrow 1990) III:124–127. *Lecture Notes in Math*, vol 1514. Springer, Berlin

81. Iwanik A (1997) Anzai skew products with Lebesgue component of infinite multiplicity. *Bull London Math Soc* 29:195–199
82. Iwanik A, Lemańczyk M, Rudolph D (1993) Absolutely continuous cocycles over irrational rotations. *Isr J Math* 83:73–95
83. Iwanik A, Lemańczyk M, de Sam Lazaro J, de la Rue T (1997) Quelques remarques sur les facteurs des systèmes dynamiques gaussiens. *Stud Math* 125:247–254
84. Iwanik A, Lemańczyk M, Mauduit C (1999) Piecewise absolutely continuous cocycles over irrational rotations. *J London Math Soc* 59(2):171–187
85. Iwanik A, de Sam Lazaro J (1991) Sur la multiplicité  $L^p$  d'un automorphisme gaussien. *CR Acad Sci Paris Ser I* 312:875–876
86. Janicki A, Weron A (1994) Simulation and chaotic behavior of  $\alpha$ -stable stochastic processes. *Monographs and Textbooks in Pure and Applied Mathematics*, vol 178. Marcel Dekker, New York, pp 355
87. del Junco A (1976) Transformations with discrete spectrum are stacking transformations. *Canad J Math* 28:836–839
88. del Junco A (1977) A transformation with simple spectrum which is not of rank one. *Canad J Math* 29:655–663
89. del Junco A (1981) Disjointness of measure-preserving transformations, minimal self-joinings and category. *Prog Math* 10:81–89
90. del Junco A, Lemańczyk M (1992) Generic spectral properties of measure-preserving maps and applications. *Proc Amer Math Soc* 115:725–736
91. del Junco A, Lemańczyk M (1999) Simple systems are disjoint with Gaussian systems. *Stud Math* 133:249–256
92. del Junco A, Lemańczyk M (2005) Joinings of distally simple systems. (in press)
93. del Junco A, Rudolph D (1987) On ergodic actions whose self-joinings are graphs. *Ergod Theory Dyn Syst* 7:531–557
94. Kachurovskii AG (1990) A property of operators generated by ergodic automorphisms. *Optimizatsiya* 47(64):122–125, (Russian)
95. Kalikow S (1984) Two fold mixing implies three fold mixing for rank one transformations. *Ergod Theory Dyn Syst* 4:237–259
96. Kamae T Spectral properties of automata generating sequences. (in press)
97. Kamiński B (1981) The theory of invariant partitions for  $\mathbb{Z}^d$ -actions. *Bull Acad Polon Sci Ser Sci Math* 29:349–362
98. Katok AB Constructions in ergodic theory. (in press)
99. Katok AB (1977) Monotone equivalence in ergodic theory. *Izv Akad Nauk SSSR Ser Mat* 41:104–157, (Russian)
100. Katok AB (1980) Interval exchange transformations and some special flows are not mixing. *Isr J Math* 35:301–310
101. Katok AB (2001) Cocycles, cohomology and combinatorial constructions in ergodic theory. In: Robinson Jr. EA (ed) *Proc Sympos Pure Math* 69, Smooth ergodic theory and its applications, Seattle, 1999. Amer Math Soc, Providence, pp 107–173
102. Katok AB (2003) Combinatorial constructions in ergodic theory and dynamics. In: *University Lecture Series* 30. Amer Math Soc, Providence
103. Katok AB, Lemańczyk M (2007) Some new cases of realization of spectral multiplicity function for ergodic transformations. (in press)
104. Katok AB, Stepin AM (1967) Approximations in ergodic theory. *Uspekhi Mat Nauk* 22(137):81–106, (Russian)
105. Katok AB, Thouvenot J-P (2006) Spectral Properties and Combinatorial Constructions in Ergodic Theory. In: *Handbook of dynamical systems*, vol 1B. Elsevier, Amsterdam, pp 649–743
106. Katok AB, Zemlyakov AN (1975) Topological transitivity of billiards in polygons. *Mat Zametki* 18:291–300, (Russian)
107. Keane M (1968) Generalized Morse sequences. *Z Wahr Verw Geb* 10:335–353
108. Keane M (1975) Interval exchange transformations. *Math Z* 141:25–31
109. Khanin KM, Sinai YG (1992) Mixing of some classes of special flows over rotations of the circle. *Funct Anal Appl* 26:155–169
110. King JL (1986) The commutant is the weak closure of the powers, for rank-1 transformations. *Ergod Theory Dyn Syst* 6:363–384
111. King JL (1988) Joining-rank and the structure of finite rank rank mixing transformations. *J Anal Math* 51:182–227
112. King JL (2001) Flat stacks, joining-closure and genericity. (in press)
113. Klemes I (1996) The spectral type of the staircase transformation. *Tohoku Math J* 48:247–248
114. Klemes I, Reinhold K (1997) Rank one transformations with singular spectral type. *Isr J Math* 98:1–14
115. Kochergin AV (1972) On the absence of mixing in special flows over the rotation of a circle and in flows on a two-dimensional torus. *Dokl Akad Nauk SSSR* 205:949–952, (Russian)
116. Kochergin AV (1975) Mixing in special flows over rearrangement of segments and in smooth flows on surfaces. *Math USSR Acad Sci* 25:441–469
117. Kochergin AV (1976) Non-degenerate saddles and absence of mixing. *Mat Zametki* 19:453–468, (Russian)
118. Kochergin AV (1976) On the homology of function over dynamical systems. *Dokl Akad Nauk SSSR* 231:795–798
119. Kochergin AV (2002) A mixing special flow over a rotation of the circle with an almost Lipschitz function. *Sb Math* 193:359–385
120. Kochergin AV (2004) Nondegenerate fixed points and mixing in flows on a two-dimensional torus II. *Math Sb* 195:15–46, (Russian)
121. Koopman BO (1931) Hamiltonian systems and transformations in Hilbert space. *Proc Nat Acad Sci USA* 17:315–318
122. Kuipers L, Niederreiter H (1974) Uniform distribution of sequences. Wiley, London, pp 407
123. Kushnirenko AG (1974) Spectral properties of some dynamical systems with polynomial divergence of orbits. *Vestnik Moskovskogo Univ* 1–3:101–108
124. Kwiatkowski J (1981) Spectral isomorphism of Morse dynamical systems *Bull Acad Polon Sci Ser Sci Math* 29:105–114
125. Kwiatkowski Jr. J, Lemańczyk M (1995) On the multiplicity function of ergodic group extensions II. *Stud Math* 116:207–215
126. Lemańczyk M (1988) Toeplitz  $Z_2$ -extensions. *Ann Inst H Poincaré* 24:1–43
127. Lemańczyk M (1996) Introduction to ergodic theory from the point of view of the spectral theory. In: *Lecture Notes of the tenth KAIST mathematics workshop*, Taejon, 1996, pp 1–153
128. Lemańczyk M (2000) Sur l'absence de mélange pour des flots spéciaux au dessus d'une rotation irrationnelle. *Coll Math* 84/85:29–41
129. Lemańczyk M, Lesigne E (2001) Ergodicity of Rokhlin cocycles. *J Anal Math* 85:43–86



130. Lemańczyk M, Mauduit C (1994) Ergodicity of a class of cocycles over irrational rotations. *J London Math Soc* 49:124–132
131. Lemańczyk M, Mentzen MK, Nakada H (2003) Semisimple extensions of irrational rotations. *Stud Math* 156:31–57
132. Lemańczyk M, Parreau F (2003) Rokhlin extensions and lifting disjointness. *Ergod Theory Dyn Syst* 23:1525–1550
133. Lemańczyk M, Parreau F (2004) Lifting mixing properties by Rokhlin cocycles. (in press)
134. Lemańczyk M, Parreau F (2007) Special flows over irrational rotation with simple convolution property. (in press)
135. Lemańczyk M, Parreau F, Roy E (2007) Systems with simple convolutions, distal simplicity and disjointness with infinitely divisible systems. (in press)
136. Lemańczyk M, Parreau F, Thouvenot J-P (2000) Gaussian automorphisms whose ergodic self-joinings are Gaussian. *Fundam Math* 164:253–293
137. Lemańczyk M, de Sam Lazaro J (1997) Spectral analysis of certain compact factors for Gaussian dynamical systems. *Isr J Math* 98:307–328
138. Lemańczyk M, Sikorski A (1987) A class of not local rank one automorphisms arising from continuous substitutions. *Prob Theory Rel Fields* 76:421–428
139. Lemańczyk M, Wasieczko M (2006) A new proof of Alexeyev's theorem. (in press)
140. Lemańczyk M, Wysokińska M (2007) On analytic flows on the torus which are disjoint from systems of probabilistic origin. *Fundam Math* 195:97–124
141. Leonov VP (1960) The use of the characteristic functional and semi-invariants in the ergodic theory of stationary processes. *Dokl Akad Nauk SSSR* 133:523–526, (1960) *Sov Math* 1:878–881
142. Maruyama G (1970) Infinitely divisible processes. *Theory Prob Appl* 15(1):1–22
143. Masur H (1982) Interval exchange transformations and measured foliations. *Ann Math* 115:169–200
144. Mathew J, Nadkarni MG (1984) Measure-preserving transformation whose spectrum has Lebesgue component of multiplicity two. *Bull London Math Soc* 16:402–406
145. Medina H (1994) Spectral types of unitary operators arising from irrational rotations on the circle group. *Michigan Math J* 41:39–49
146. Mentzen MK (1988) Some examples of automorphisms with rank  $r$  and simple spectrum. *Bull Pol Acad Sci* 7–8:417–424
147. Nadkarni MG (1998) Spectral theory of dynamical systems. Hindustan Book Agency, New Delhi
148. von Neumann J (1932) Zur Operatorenmethode in der Klassischen Mechanik. *Ann Math* 33:587–642
149. Newton D (1966) On Gaussian processes with simple spectrum. *Z Wahrscheinlichkeitstheorie Verw Geb* 5:207–209
150. Newton D, Parry W (1966) On a factor automorphism of a normal dynamical system. *Ann Math Statist* 37:1528–1533
151. Ornstein D (1970) On the root problem in ergodic theory. In: *Proc. 6th Berkeley Symp Math Stats Prob*, University California Press, Berkeley, 1970, pp 348–356
152. Ornstein D, Rudolph D, Weiss B (1982) Equivalence of measure preserving transformations. *Mem Amer Math Soc* 37(262):116
153. Ornstein D, Weiss B (1987) Entropy and isomorphism theorems for actions of amenable groups. *J Anal Math* 48:1–141
154. Parreau F (2000) On the Foias and Stratila theorem. In: *Proceedings of the conference on Ergodic Theory*, Toruń, 2000
155. Parreau F, Roy E (2007) Poisson suspensions with a minimal set of self-joinings. (in press)
156. Parry W (1981) Topics in ergodic theory. *Cambridge Tracts in Mathematics*, 75. Cambridge University Press, Cambridge-New York
157. Parry W (1970) Spectral analysis of  $G$ -extensions of dynamical systems. *Topology* 9:217–224
158. Petersen K (1983) Ergodic theory. Cambridge University Press, Cambridge
159. Prikhodko AA, Ryzhikov VV (2000) Disjointness of the convolutions for Chacon's automorphism. Dedicated to the memory of Anzelm Iwanik. *Colloq Math* 84/85:67–74
160. Queffelec M (1988) Substitution dynamical systems – spectral analysis. In: *Lecture Notes in Math* 1294. Springer, Berlin, pp 240
161. Ratner M (1978) Horocycle flows are loosely Bernoulli. *Isr J Math* 31:122–132
162. Ratner M (1983) Horocycle flows, joinings and rigidity of products. *Ann Math* 118:277–313
163. Rauzy G (1979) Echanges d'intervalles et transformations induites. *Acta Arith* 34:315–328
164. Robinson EA Jr (1983) Ergodic measure preserving transformations with arbitrary finite spectral multiplicities. *Invent Math* 72:299–314
165. Robinson EA Jr (1986) Transformations with highly nonhomogeneous spectrum of finite multiplicity. *Isr J Math* 56:75–88
166. Robinson EA Jr (1988) Nonabelian extensions have nonsimple spectrum. *Compos Math* 65:155–170
167. Robinson EA Jr (1992) A general condition for lifting theorems. *Trans Amer Math Soc* 330:725–755
168. Rosiński J, Zak T (1996) Simple condition for mixing of infinitely divisible processes. *Stoch Process Appl* 61:277–288
169. Rosiński J, Zak T (1997) The equivalence of ergodicity and weak mixing for infinitely divisible processes. *J Theor Probab* 10:73–86
170. Roy E (2005) Mesures de Poisson, infinie divisibilité et propriétés ergodiques. In: *Thèse de doctorat de l'Université Paris 6*
171. Roy E (2007) Ergodic properties of Poissonian ID processes. *Ann Probab* 35:551–576
172. Royden HL (1968) Real analysis. McMillan, New York
173. Rudin W (1962) Fourier analysis on groups. In: *Interscience Tracts in Pure and Applied Mathematics*, No. 12 Interscience Publishers. Wiley, New York, London, pp 285
174. Rudolph DJ (1979) An example of a measure-preserving map with minimal self-joinings and applications. *J Anal Math* 35:97–122
175. Rudolph D (1985)  $k$ -fold mixing lifts to weakly mixing isometric extensions. *Ergod Theor Dyn Syst* 5:445–447
176. Rudolph D (1986)  $\mathbb{Z}^n$  and  $\mathbb{R}^n$  cocycle extension and complementary algebras. *Ergod Theor Dyn Syst* 6:583–599
177. Rudolph D (1990) Fundamentals of measurable dynamics. Oxford Sci Publ, pp 169
178. Rudolph D (2004) Pointwise and  $L^1$  mixing relative to a sub-sigma algebra. *Illinois J Math* 48:505–517
179. Rudolph D, Weiss B (2000) Entropy and mixing for amenable group actions. *Ann Math* 151(2):1119–1150
180. de la Rue T (1996) Systèmes dynamiques gaussiens d'entropie nulle, lâchement et non lâchement Bernoulli. *Ergod Theor Dyn Syst* 16:379–404

181. de la Rue T (1998) Rang des systèmes dynamiques gaussiens. *Isr J Math* 104:261–283
182. de la Rue T (1998) L'ergodicité induit un type spectral maximal équivalent à la mesure de Lebesgue. *Ann Inst H Poincaré Probab Statist* 34:249–263
183. de la Rue T (1998) L'induction ne donne pas toutes les mesures spectrales. *Ergod Theor Dyn Syst* 18:1447–1466
184. de la Rue T (2004) An extension which is relatively twofold mixing but not threefold mixing. *Colloq Math* 101:271–277
185. Ryzhikov VV (1991) Joinings of dynamical systems. Approximations and mixing. *Uspekhi Mat Nauk* 46 5(281):177–178 (in Russian); *Math Surv* 46(5):199–200
186. Ryzhikov VV (1992) Mixing, rank and minimal self-joining of actions with invariant measure. *Math Sb* 183:133–160, (Russian)
187. Ryzhikov VV (1994) The absence of mixing in special flows over rearrangements of segments. *Math Zametki* 55:146–149, (Russian)
188. Ryzhikov VV (1999) Transformations having homogeneous spectra. *J Dyn Control Syst* 5:145–148
189. Ryzhikov VV (2000) The Rokhlin problem on multiple mixing in the class of actions of positive local rank. *Funkt. Anal Prilozhen* 34:90–93, (Russian)
190. Ryzhikov VV, Thouvenot J-P (2006) Disjointness, divisibility, and quasi-simplicity of measure-preserving actions. *Funkt. Anal Prilozhen* 40:85–89, (Russian)
191. Ryzhikov VV (2007) Weak limits of powers, simple spectrum symmetric products and mixing rank one constructions. *Math Sb* 198:137–159
192. Sato K-I (1999) Lévy Processes and infinitely divisible distributions. In: *Cambridge Studies in Advanced Mathematics*, vol 68. Cambridge University Press, Cambridge, pp 486
193. Schmidt K (1977) Cocycles of Ergodic Transformation Groups. In: *Lecture Notes in Math.* 1. Mac Millan, India
194. Schmidt K (2002) Dispersing cocycles and mixing flows under functions. *Fund Math* 173:191–199
195. Schmidt K, Walters P (1982) Mildly mixing actions of locally compact groups. *Proc London Math Soc* 45(3):506–518
196. Sinai YG (1994) Topics in ergodic theory. Princeton University Press, Princeton
197. Shklover M (1967) Classical dynamical systems on the torus with continuous spectrum. *Izv Vys Uchebn Zaved Matematika* 10:(65)113–124, (Russian)
198. Smorodinsky M, Thouvenot J-P (1979) Bernoulli factors that span a transformation. *Isr J Math* 32:39–43
199. Stepin AM (1986) Spectral properties of generic dynamical systems. *Izv Akad Nauk SSSR Ser Mat* 50:801–834, (Russian)
200. Tikhonov SV (2002) On a relation between the metric and spectral properties of  $\mathbb{Z}^d$ -actions. *Fundam Prikl Mat* 8:1179–1192
201. Thouvenot J-P (1995) Some properties and applications of joinings in ergodic theory. *London Math Soc Lect Note Ser* 205:207–235
202. Thouvenot J-P (2000) Les systèmes simples sont disjoints de ceux qui sont infiniment divisibles et plongeables dans un flot. *Coll Math* 84/85:481–483
203. Ulcigrai C (2007) Mixing of asymmetric logarithmic suspension flows over interval exchange transformations. *Ergod Theor Dyn Syst* 27:991–1035
204. Walters P (1982) An Introduction to Ergodic Theory. In: *Graduate Texts in Mathematics*, vol 79. Springer, Berlin, pp 250
205. Wysokińska M (2004) A class of real cocycles over an irrational rotation for which Rokhlin cocycle extensions have Lebesgue component in the spectrum. *Topol Meth Nonlinear Anal* 24:387–407
206. Wysokińska M (2007) Ergodic properties of skew products and analytic special flows over rotations. In: Ph D thesis. Toruń, 2007.
207. Veech WA (1978) Interval exchange transformations. *J Anal Math* 33:222–272
208. Veech WA (1982) A criterion for a process to be prime. *Monatshfte Math* 94:335–341
209. Veech WA (1982) Gauss measures for transformations on the space of interval exchange maps. *Ann Math* 115(2):201–242
210. Veech WA (1984) The metric theory of interval exchange transformations, I Generic spectral properties. *Amer J Math* 106:1331–1359
211. Vershik AM (1962) On (1962) the theory of normal dynamic systems. *Math Sov Dokl* 144:625–628
212. Vershik AM (1962) Spectral and metric isomorphism of some normal dynamical systems. *Math Sov Dokl* 144:693–696
213. Yassawi R (2003) Multiple mixing and local rank group actions. *Ergod Theor Dyn Syst* 23:1275–1304
214. Zeitz P (1993) The centralizer of a rank one flow. *Isr J Math* 84:129–145

## Spin Dependent Exchange and Correlation in Two-Dimensional Electron Layers

M. W. CHANDRE DHARMA-WARDANA

Institute of Microstructural Sciences,  
National Research Council of Canada, Ottawa, Canada

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Ideally Thin 2D Electron Layers](#)

[2D Layers with Finite Thickness](#)

[Multi-Component 2D Layers](#)

[Graphene: 2D Two-Valley System](#)

[on a Honey-Comb Lattice](#)

[Exchange and Correlation in a Magnetic Field](#)

[Exchange and Correlation at Finite Temperatures](#)

[Conclusion](#)

[Bibliography](#)

### Glossary

**Atomic units, a. u.** The electron-charge  $|e|$ , and the mass  $m_e$  are taken as unity. The unit of time is fixed by setting the Plank constant  $\hbar$  to unity. The Bohr radius  $a_0 = \hbar^2/(m_e e^2)$  is one a.u. of length in the

centimeter-gram-second (CGS) system which uses the “esu” (electrostatic unit of charge). The SI system uses the meter, kilogram, second (with the Ampere as the unit of current),  $a_0 = (4\pi\epsilon_0\hbar^2/(m_e e^2))$  where  $\epsilon$  is the electric permittivity of the vacuum. The value of  $a_0$  is  $5.29177 \times 10^{-9}$  cm. The unit of energy, the Hartree, is  $e^2/a_0$  in the CGS system, being 27.2116 eV. In semiconductor physics, *effective atomic units* are used, with  $e^2/\epsilon$ , replacing  $e^2$ , where  $\epsilon$  is the dielectric constant. The band mass  $m_b$  is used instead of  $m_e$ , so that the effective Bohr radius  $a_0^* = a_0(\epsilon/m^*)$ , where  $m^*$  is the effective mass. Then the effective Hartree is of the order of millivolts.

**Confining potential** This potential keeps the electron in a given spatial region. It is due to the physical structure of the device, the applied gate voltages etc.

**Correlation energy** The contribution to the total energy beyond the Hartree–Fock approximation, denoted by  $E_c$ .

**Correlation hole** is the depletion of electron density near an electron due to Coulomb repulsion effects.

**Coupling constant** is the ratio of the potential energy (PE) to the kinetic energy (KE). In a classical electron fluid, the KE (thermal energy) is  $T$ , and the PE is  $1/r_s$  (atomic units), and  $r_s$  is the Wigner–Seitz radius. The coupling constant = PE/KE =  $\Gamma = 1/(r_s T)$ . In quantum systems, the Fermi energy  $E_F$  is used instead of  $T$  for the KE, and  $\Gamma = r_s$ .

**Classical-map hyper-netted-chain (CHNC)** A method for using the classical hyper-netted-chain equation to calculate the correlation functions of quantum systems.

**Effective mass** is denoted by  $m^*$ , and is the band mass  $m_b$  in units of the electron mass  $m_e$ .

**Exchange energy** The part of the Hartree–Fock energy due to electron exchange, i. e., the “Fock” part, denoted by  $E_x$ . It is first order in the Coulomb interaction.

**Fermi hole** This denotes the reduction in the probability of finding a like-spin electron near another, due to Fermi statistics.

**Hartree–Fock** Hartree’s self-consistent one-body approximation for interacting electrons is based on a product wavefunction. Fock included exchange using an antisymmetrized product. “Hartree–Fock” is the label for calculations of the energy, wavefunctions etc., where the electron moves in this mean potential generated by the electrostatics and the exchange effects. The Hartree term is zero in uniform systems.

## HIGFET

Heterojunction-Insulated-Gate Field-effect Transistor.

**Jellium** A model “metal” where the positive ionic charges are replaced by a uniform static charge which neutralizes the free-electron charge.

**MOSFET** Metal-oxide semiconductor field-effect transistor.

**Heterojunction** A semiconductor-interface involving two dissimilar materials.

**Hyper-netted-chain (HNC)** A classical integral equation due to van Leeuwen, Groenvelde and de Boer (1959) which non-perturbatively sums “hyper-netted-chain” diagrams, going beyond mean-field theory.

**Electron gas parameter  $r_s$**  See Wigner–Seitz radius.

**Pair-correlation function** Denoted by  $h(\vec{r}) = g(\vec{r}) - 1$  where  $g(r)$  is the pair-distribution function (PDF).

**Pair-distribution function** The pair-distribution function (PDF),  $g(\vec{r})$ , is the probability of finding a particle at the location  $\vec{r}$ , given a reference particle at the origin.

**Plasma analogy** A class of methods for approximately replacing a charged quantum fluid by an equivalent classical fluid at a finite temperature.

**Pseudo-spin** Discrete degrees of freedom beside the electron spin. The electrons in Si/SiO<sub>2</sub> interfaces occupy two valleys. The valley index is a pseudospin.

**Quantum-Monte Carlo (QMC)** In molecular dynamics (MD), Newton’s equations of motion are integrated using a stochastic scheme based on the Metropolis algorithm. In QMC a trial wavefunction provides a probability measure for the Metropolis algorithm. The wavefunction is optimized in various ways, leading to “variational QMC”, where the nodes of the trial wavefunction are held fixed. In “Diffusion QMC”, the nodes are also relaxed.

**Random-phase approximation (RPA)** A time-dependent self-consistent field method where an electron with momentum  $\vec{k}$  moves in an effective potential which contains the external potential and a  $\vec{k}, \omega$  dependent screened potential. It reduces in the static  $k \rightarrow 0$  limit to Thomas–Fermi screening, (or Debye–Hückel screening in classical systems). It is also called the “ring sum” or “bubble sum”, and contains no exchange effects.

**Subbands** The electrons with the  $z$ -motion confined to a quantum well have discrete energy levels (index  $n$ ). Each level carries with it a band of energies for the  $x, y$  motion. These are energy “subbands”.

**Singwi–Tosi–Land–Sjölander (STLS)** A method due to Singwi et al. for determining the density-density correlation function of electrons (and other quantum systems) by truncating the equation of motion via an intuitive decoupling scheme involving the PDFs. STLS has been extended by Vashista, Ichimaru and others.

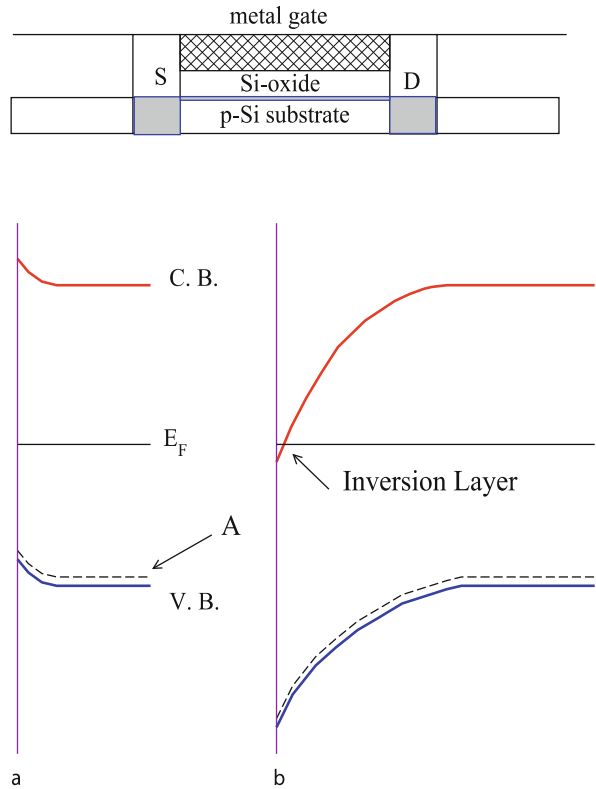
**Wigner–Seitz radius** In 2D this is the radius, denoted by  $r_s$ , of the circle containing, on the average, just one electron.

### Definition of the Subject

Since the advent of density-functional theory (DFT), the exchange-correlation energy  $E_{xc}$  of an interacting system has become a basic quantity in many-particle theory. Here we study the  $E_{xc}$  of two-dimensional (2D) electron layers. Such layers contain electrons which move in the  $x$  and  $y$  directions, while confined in  $z$ . 2D layers are formed at insulator-semiconductor interfaces in heterojunctions, and more particularly at metal-oxide-semiconductor (MOS) interfaces. These include two types of semiconductors (e.g., GaAs and the alloyed form  $\text{Al}_x\text{Ga}_{1-x}\text{As}$ , containing a small fraction  $x$  of Al, and written as AlGaAs for brevity). The interface region defines a “confining potential” where an electron layer may form [2].  $\text{SiO}_2$  is an insulator with a large bandgap, while Si can be doped in a controlled manner to behave as a conductor. The Si/SiO<sub>2</sub> interface supports the formation of an electron layer at the interface. The electron density  $n$  in such layers can be controlled using external potentials, and it is this which is the key to the importance of these materials systems. These structures are the basis of metal-oxide field-effect transistors (MOSFETs), ubiquitous in modern electronic devices (Fig. 1). The Si-MOSFET was developed in the 1960s, while it had been fore-shadowed as early as the 1930s in the work of Shockley and others.

Electron layers confined at the air-liquid interface of a fluid (e.g., Helium) were studied in the 1970s, but did not provide easy tunability of the density. High-mobility 2D systems in GaAs/GaAlAs heterostructures and in Si-MOSFETs ushered the more recent phase, unearthing new physics and new technologies, with the quantum Hall effect (QHE) [4], discovered in 1980, providing a classic example of fundamental physics directly leading to practical applications in metrology. Modern nanotechnology, spintronics, plasmonics etc., greatly dependent on the theory of electron layers.

Graphite is made of 2D sheets of carbon atoms, i.e., “graphene” sheets, held together by weak inter-layer bonds. The 2D electrons in graphene behave as massless Fermions with  $\epsilon(\vec{k}) = \hbar v_F \vec{k}$ , where  $v_F$  is the Fermi velocity. The hexagonal unit cell of graphene contains two inequivalent Carbon atoms. Thus there are two degenerate conduction bands and two valance bands, with a zero band gap located at two inequivalent points, labeled  $\mathbf{K}$  and  $\mathbf{K}'$ , in the hexagonal 2D-Brillouin zone of graphene. This system needs a spin index, a pseudospin index and also a band



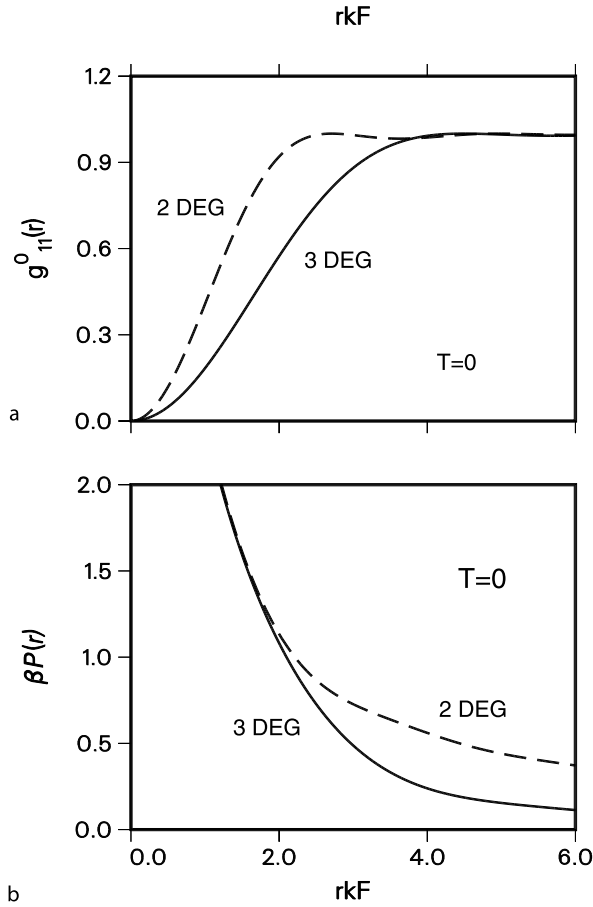
**Spin Dependent Exchange and Correlation in Two-Dimensional Electron Layers, Figure 1**

**Top:** A cross section of a p-type MOSFET. The n-type contacts labeled S (source) and D (drain) are made by ion-implantation into the p-type substrate. A voltage on the metal gate controls the current between S and D flowing in the Si/SiO<sub>2</sub> interface (channel region). **Panel a** shows the conduction band (c.b.), the acceptor levels (A), and the valance band (v.b.) modified by the application of negative bias to the metal gate. A sufficiently strong negative bias can produce an accumulation layer of holes. In **b** a sufficiently positive gate voltage is applied, bending the conduction band *below* the Fermi energy  $E_F$ , and creating a potential well which supports a 2D layer of electrons, known as an ‘inversion layer’

index. Graphene is a rich 2D system with novel physics (e.g. QHE) and the promise of new technological applications [9].

The Coulomb interaction conspires to destroy the “single-particle” (i.e., “free” particle) picture of electron systems. Only the kinetic energy and the confining potential (called the “external potential”) are needed in the simplest models. Mean-field (e.g. Hartree–Fock) models provide an “effective” single-particle picture. The “many-body effects”, which transcend the simple picture, lead to new effects like plasmons, charge density waves, spin-polarized states, Wigner crystallization, fractional-QHE, and super-





**Spin Dependent Exchange and Correlation in Two-Dimensional Electron Layers, Figure 2**

Upper panel shows the pair-distribution function (PDF) for non-interacting parallel-spin electrons in 2D and 3D electron gases. The depletion near the origin is the "Fermi hole". The lower panel shows the equivalent classical potential (Pauli exclusion potential, in  $\beta = 1/T$  units) which would create such a depletion in the PDF at the temperature  $T$  (at  $T = 0$ , only the product  $\beta P(r)$  is meaningful). The x-axis is the electron-pair separation in reciprocal  $k_F$  units

conductivity [18]. Many-body effects are more enhanced in low dimensional (e. g., 1D and 2D systems in comparison to 3D) systems. The Coulomb interactions force two electrons to repel each other and form a "Coulomb hole" or correlation hole, i. e., a region of low probability of approach. The energy consequence of this is the "correlation energy". Similarly, since the Pauli principle forbids the presence of two non-interacting electrons of the same spin in a common spatial eigenfunction, a "Fermi hole" is formed and leads to an exchange energy (see Fig. 2).

There is no Fermi hole for antiparallel spins, and their non-interacting PDF, i. e.,  $g_{12}^0(r)$  is unity for all  $r$ . When in-

teractions are included, a Coulomb hole is formed. Hence the evaluation of  $E_{xc}$  is directly related to obtaining good PDFs.

Thus the theory of  $E_{xc}$  is of great importance in modern density-functional many-body approaches. From a technological point of view, the subject is of critical importance in the proper design of modern nano-structure devices and quantum-well lasers [16].

## Introduction

We use "effective" atomic units, with the effective Bohr radius  $a_0^*$  as the unit of length. The dielectric constant  $\epsilon$  is of the order of 10–12 for common semiconductors, while  $m^*$  may be 0.06 in GaAs and 0.19 in Si. Hence the effective Hartree is measured in meV, unlike the atomic Hartree which is 27.12 eV. Thus the properties of the material enter into the theory only via  $\epsilon$ ,  $m^*$ , and the effective Landé factor  $g^*$ , and these are usually absorbed into the effective atomic units. Since the electrons reside in the interface between two materials (say A, B), the dielectric constant  $\epsilon$  relevant to the electron layer has to be evaluated from the individual dielectric constants of A and B. This is trivial in the case of GaAs/AlGaAs systems, since the dielectric constants are very close. This is no longer the case with Si/SiO<sub>2</sub> systems where the Si and the Si-oxide have nominal dielectric constants of 11.5 and 3.9 respectively. Then an average dielectric constant has to be used, as discussed in the appendix of [2], and in [8].

If the  $z$ -motion of an electron is restricted to a finite length  $a_z$ , as in an inversion layer, the allowed quantum states form discrete subbands, with energies  $E(n, k_x, k_y)$ :

$$E_n(k_x, k_y) = \epsilon_n + \epsilon(k_x, k_y).$$

The energy of the in-plane motion,  $\epsilon(k_x, k_y)$ , depends on the in-plane momenta  $\hbar k_x$  and  $\hbar k_y$  with  $\vec{k} = \vec{k}_x + \vec{k}_y$ . For materials like GaAs/AlGaAs, or Si/SiO<sub>2</sub> interfaces, the in-plane energy dispersion is essentially parabolic, with

$$\epsilon(k_x, k_y) = \frac{(\hbar k_x)^2 + (\hbar k_y)^2}{2m^*}.$$

Here  $m^*$  is an effective mass, assumed to be the same in the  $x$  and  $y$  directions. Only the lowest subband, say  $n = 1$ , is occupied in a 2D layer. Let  $\epsilon_1$  be the energy zero. The highest occupied state is given by the in-plane momentum  $\hbar k_F$ , known as the *Fermi momentum*. The Fermi energy ( $\epsilon_{k_F}$ ) has to be far below the bottom of the second subband to ensure a 2D layer. The electron density, i. e., the number of electrons  $n$  per unit area has  $n_1$  "up-spin", and  $n_2$  "down-spin" electrons. If  $n_1 \neq n_2$  we have a *spin-polarized* electron system. The *degree of spin polarization*

$\zeta$  is:

$$\zeta = \frac{n_1 - n_2}{n_1 + n_2}.$$

Thus the state of a clean 2D-electron layer may be specified by  $n$ ,  $\zeta$ , and the temperature  $T$ . Additional pseudo-spin or valley indices are sometimes necessary.

### The Fermi Energy and the Role of $r_s$ as a Coupling Constant

When the electron-number density per a.u. of area is  $n$ , the Wigner-Seitz radius  $r_s$  is given by  $r_s = 1/\sqrt{(\pi n)}$ . We also have  $n_1 = nx_1$ ,  $x_1 = (1 + \zeta)/2$ ,  $n_2 = nx_2$ ,  $x_2 = (1 - \zeta)/2$ . The Fermi momentum  $k_{F\sigma}$  of a given spin species  $\sigma$  is evaluated by requiring that the sum of all occupied states of the species  $\sigma$  adds upto the density  $n_\sigma$ .

$$\int_0^{k_F} 2\pi \frac{k dk}{(2\pi)^2} = n_\sigma.$$

This gives  $k_\sigma = (4\pi n_\sigma)^{1/2}$ . Since  $n = n_1 + n_2$ , the overall  $k_F = (2\pi n)^{1/2}$ , i. e.,  $k_F = \sqrt{2}/r_s$  and  $E_F = k_F^2/2 = 1/r_s^2$ .

Hence, the “coupling strength”, i. e., ratio of the Coulomb energy ( $1/r_s$ ) to the kinetic energy ( $\sim E_F$ ) at  $T = 0$  is clearly  $r_s$ . High density systems ( $r_s < 1$ ) are weakly coupled, while low density systems (high  $r_s$ ) are strongly correlated. When  $r_s \sim 26 - 27$ , the 2D electron system acquires a ferromagnetic ground state. At stronger coupling,  $r_s \simeq 35$ , a 2D Wigner crystal is formed. The strength of the 2D massless electron interactions in graphene cannot be specified by an  $r_s$  since a Bohr radius cannot be specified. Instead, the coupling constant  $g$  is taken as the ratio of a typical Coulomb energy ( $1/a_0$ ) to the hopping energy ( $t$ ) on the Hexagonal 2D unit cell with a lattice constant of  $a_0$ .

$$g = \frac{e^2/a_0\epsilon_0}{\hbar E_F} = \{e^2/a_0\epsilon_0\} \{t\sqrt{3}/2\}.$$

If  $\epsilon_0 = 1$ , then  $g \simeq 2.7$ . Thus, when the graphene sheet is placed on most substrates,  $\epsilon > 1$ , and the electron interactions become weak, i. e.,  $g \leq 1$ .

### Fourier Transforms

The Fourier transform of 2D-functions relating their  $r$ -space forms with  $q$ -space forms is very useful. If the 2D electron system is uniform in all directions of the  $x - y$  plane, i. e., isotropic, then we only need *radial* Fourier

transforms. These are obtained using the relations:

$$F(\vec{r}) = \int_0^\infty F(\vec{q}) J_0(qr) q dq / (2\pi)$$

$$F(\vec{q}) = \int_0^\infty F(\vec{r}) J_0(qr) 2\pi r dr.$$

Here  $J_0(qr)$  is a Bessel function [1] such that:

$$J_0(qr) = 2 \int_0^\pi \cos(qr \cos \theta) d\theta.$$

Then it is easy to show that the Coulomb interaction  $1/r$  has the Fourier form  $2\pi/q$ . Here we assume ideally thin layers (unlike in Sect. “[Graphene: 2D Two-Valley System on a Honey-Comb Lattice](#)”). A screened Coulomb interaction  $V_s(r)$ , with an exponential damping is often encountered. This potential has an analytic Fourier transform and is:

$$V_s(r) = \exp(-k_s r) / r \quad V_s(q) = 2\pi / \{(q^2 + k_s^2)^{1/2}\}.$$

The potential contains a “screening wavevector”  $k_s$ , and appears in what is known as “Thomas-Fermi” screening. This type of potential is also called a “Yukawa potential”. The ‘range’ of the potential is of the order of  $1/k_s$ , and is known as the screening length.

### The Hamiltonian of the System

The total Hamiltonian contains a contribution from a *uniform* neutralizing background of positive charge  $n_b$  which is static. In a real system, e. g., a metal, the neutralizing background is provided by the ion subsystem. Each ion has a short-ranged core-region and a Coulomb-like long-range potential which overlaps the long-range potential of the neighbours. In many metals (e. g., Sodium) the resulting overlapping positive potential is, to a very good approximation, similar to a smudged-out, structureless “jellium”. In a plasma, the positive ions are dynamical but extremely heavy, and hence a jellium approximation, known as the one-component plasma (OCP) is often used. The jellium model is used in theories of electron layers as the physics then focuses entirely on electron-electron effects.

We write the electron density operator and the Coulomb interaction as:

$$n(\vec{r}) = \sum_i \delta(\vec{r} - \vec{r}_i), \quad V_{ee} = \frac{1}{2} \sum_{ij} \frac{n(\vec{r}_i) n(\vec{r}_j)}{|\vec{r}_i - \vec{r}_j|}.$$

The latter has a divergence when  $i \rightarrow j$ , with the electron interacting with “itself”. We also have the terms  $V_{bb}$  and  $V_{be}$ , involving the background. For example,

$$V_{be} = - \sum_{ij} n_b \frac{n(\vec{r}_j)}{|\vec{r}_i - \vec{r}_j|}$$

where  $n_b = n$  is the uniform-background density. The background term  $V_{bb}$  has a divergent self-interaction term similar to that in  $V_{ee}$ , with (positive) sign. However, the interaction of the electron charge with the background, i. e.,  $V_{be}$ , has a negative divergent term, and cancels both divergent (positive) self-interaction terms. Thus the total Hamiltonian, inclusive of the background is free of divergences. In Fourier space, the coulomb interaction is  $2\pi/q$ , and hence the divergence manifests in the limit  $q \rightarrow 0$ . This is exactly canceled by the corresponding  $q \rightarrow 0$  terms in the background. Also, since the background is uniform, there are no other Fourier components. The above discussion involves the “subtraction” of divergent quantities, and can be made more rigorous using the screened Coulomb potential  $V_s(r)$ , and taking the limit  $k_s \rightarrow 0$  at the end.

The electron number-density operator  $n(r)$  in real space has a Fourier transform  $n(q)$ , defining the occupation number in the momentum state  $q$ . The expectation value of  $n(r)$  is of course the constant  $n$ , which may be written as  $\bar{n}$  where needed.

The momentum eigenstates of the uniform noninteracting system are the complete set of plane waves.

$$\phi_q = \phi(\vec{q}) = e^{i\vec{q}\cdot\vec{r}}/\Omega^{1/2}$$

Here we use a normalization volume  $\Omega$ , which is sometimes set to unity, or the infinite-volume limit is taken, when the normalization becomes  $1/(2\pi)^{1/2}$ . The occupation number in each momentum state, for non-interacting Fermions at a temperature  $T$ , and chemical potential  $\mu$  is given by:

$$n_k = \frac{1}{1 + e^{\beta(\epsilon_k - \mu)}}.$$

Here  $\beta = 1/T$  where  $T$  is expressed in energy units. At  $T = 0$ , the chemical potential  $\mu$  becomes the Fermi energy  $E_F$ , and the occupation numbers reduce to unity for energies  $\epsilon_k < E_F$ , and zero for higher energies. The occupation number  $n_k$  given above is really the expectation value of the operator  $n_k$ . When we need to emphasize this distinction, we write the operator as  $\hat{n}_k$ , and the meanvalue as  $\bar{n}_k$ . In a uniform system, most properties are dependent only on the modulus  $|k|$  of the momentum  $\vec{k}$ , and we sometimes suppress the vector notation for brevity. We

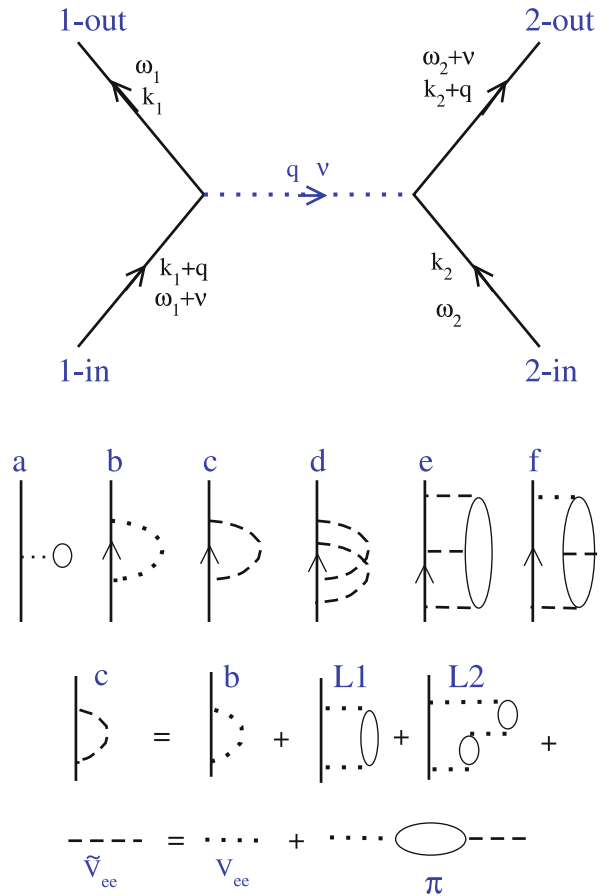
also introduce creation and annihilation operators  $a_k^\dagger$  and  $a_k$  which add an electron to a momentum state  $\vec{k}$ , or remove an electron from the momentum state  $\vec{k}$ . Thus the number operator acting on the eigenstate  $|\phi_k\rangle$  is given as:

$$\hat{n}_k |\phi_k\rangle = a_k^\dagger a_k |\phi_k\rangle = \bar{n}_k |\phi_k\rangle.$$

The total Hamiltonian of the system is:

$$H = (H^0 - \mu N) + V'_{ee}.$$

Here we are using a thermal ensemble which allows for the definition of a chemical potential  $\mu$  and a temperature  $T$ . Also,  $N = \sum_k \hat{n}_k$  is an operator fixing the mean



**Spin Dependent Exchange and Correlation in Two-Dimensional Electron Layers, Figure 3**

**Top Panel:** Feynman diagram for the bare Coulomb interaction  $V_{ee}(q)$  of two electrons. **Bottom panel:** **a** shows the exchange interaction which is first order in the bare interaction  $V_{ee}(q)$ . **b** is the bare exchange, and **c** has the RPA “bubble sum”, where  $\tilde{V}_{ee}(q, v)$  is given by an integral equation (a geometric series in this case) involving the polarization “bubble”  $\pi(q, v)$ . Contributions from diagrams like **c–d** are not included in the RPA propagators

number of particles  $\langle N \rangle$  in the volume  $\Omega$ , given  $\mu$  and  $T$ . The Coulomb interaction  $V'_{ee}$  carries a prime to indicate that a jellium background has been included to remove the divergencies at zero momentum transfer. Then we have, in momentum space:

$$H = \sum_{\vec{k}} (\epsilon_k - \mu) a_k^+ a_k + \frac{1}{2\Omega} \sum_{\vec{k}_1, \vec{k}_2, \vec{q}} V_q a_{\vec{k}_1}^+ a_{\vec{k}_2 + \vec{q}}^+ a_{\vec{k}_2} a_{\vec{k}_1 + \vec{q}}.$$

The prime on the summation indicates that the case  $q = 0$  is excluded, as it is exactly cancelled by the jellium background. The suffixes on the cluster of four operators are such that the total momentum of the two electrons is conserved during the interaction. That is, an electron in the initial state  $\vec{k}_1 + \vec{q}$  is in the final state  $\vec{k}_1$ , while the second electron, in the initial state  $\vec{k}_2$  is transferred to the state  $\vec{k}_2 + \vec{q}$ . The momentum transfer  $\vec{q}$  is effected through the Coulomb interaction  $V_q = 2\pi/q$ . The momentum is conserved in the Coulomb collision since we are dealing with a *uniform system* (see top panel in Fig. 3).

### Ideally Thin 2D Electron Layers

This system is most closely approximated in Si/SiO<sub>2</sub> systems, or in specially fabricated GaAs/GaAlAs quantum wells (HIGFETs tend to yield thick 2D electron layers and are discussed in Sect. “2D Layers with Finite Thickness”). Microscopic many-body theories have used this system as the basic “work horse”. Early work used diagrammatic and other perturbation methods, or truncated equation of motion methods, like STLS, developed by Singwi and coworkers. The objective of the latter is to develop PDFs non-perturbatively, and calculate the exchange and correlation energies from an integration over the coupling constant (described below). However, the perturbation methods (restricted to the high-density regime in validity) provide basic reference results.

### A Brief Outline of the Diagrammatic Method

Much of the initial work was based on a perturbation expansion of the energy or the electron propagator in powers of the Coulomb interaction  $V_{ee}(q)$ , as in Fig. 3. In the graphs labeled a–f we look at the self-energy (i.e., the effective potential) of an electron propagating through the interacting fluid. In (a) an electron propagates forward, while the interaction with the second electron occurs with zero-momentum transfer ( $q = 0$ ), i.e., one may imagine that the “2-out” line “falls into” the “2-in” line to form a bubble, when  $\vec{k}_2 + \vec{q} = \vec{k}_2$  by momentum conservation. Such  $q = 0$  terms are Hartree terms, and have

already been eliminated via the static background. The graph (b) is the first-order exchange diagram (the “Fock” term). This may be pictured as the “1-out” line “falling” into the “2-in” line, exchanging electron identities. This is possible only if both electrons have the same spin. The graph (c) is the *screened exchange* diagram where the Coulomb interaction is *dynamically screened* by summing the simple polarization loops L1, L2 etc. This is known as the RPA-screened exchange, where RPA (random-phase approximation) is a name originating from methods of derivation using self-consistent equations. However, each term in such an expansion is infinite; e.g., the second-order perturbation term L1, and those beyond it (L2 etc.) are all divergent, as they involve integrals of the form  $\int (2\pi/q)^n q dq \dots$ , where  $n \geq 2$  (there are standard “Feynman rules” for converting the graphs shown in Fig. 3 into algebraic expressions [18]; they are not needed for our purpose). The divergence arises from perturbation terms containing simple polarization loops (denoted by  $\pi(q, \omega)$  in Fig. 3) connected by two or more direct interaction lines. However, although each term is infinite, the *sum* L1 + L2 + ... is finite. The bare interaction  $V_{ee}(q, \omega)$  gets “screened” by the polarization processes denoted by the bubble (particle-hole pairs), and gives the well-behaved (non-divergent)  $\tilde{V}_{ee}(q)$  which is used in the graph (c). The contributions from scattering processes shown in Fig. 3d–f, etc., are *not* included in the RPA sum.

Thus the contribution to the effective potential seen by an electron at the Fermi level can be written as

$$V_{\text{eff}} = V_{\text{xc}} = V_x + V_c$$

where  $V_x$  is the exchange potential, arising from the first-order diagram Fig. 3b, while  $V_c$  is the contribution from *all other* higher processes. Thus  $V_c$  in the RPA is simply the sum of contributions from L1, L2 etc. A better approximation would be a  $V_c$  which involves Fig. 3e–f, and all other terms. But the numerous higher terms are virtually impossible to evaluate, and their partial inclusion is problematic as various sum rules, Ward identities etc., have to be satisfied [18]. In analogy with the effective potential, we can also divide the total energy of the system (beyond the kinetic energy) as:

$$E_{\text{xc}} = E_x + E_c; \quad \epsilon_x = E_x/N; \quad \epsilon_c = E_c/N$$

where  $E_x$  is the exchange energy, and  $E_c$  is the correlation energy. The latter by definition contains all contributions other than exchange. A large part of the spin dependence of  $E_{\text{xc}}$  is contained in  $E_x$ . The RPA is an evaluation of  $E_c$  via the series L1, L2 etc.



### The Adiabatic Connection Formula

Another approach to  $E_{xc}$  is to begin with the non-interacting system and “build-up” the interacting system by increasing the value of the “coupling constant”  $\lambda$  from  $0 \rightarrow 1$  in the scaled Coulomb potential

$$V_{ee}(\lambda, r) = \lambda/r.$$

The usual coupling parameter  $r_s$  is here assumed to be merely a density parameter ( $r_s = 1/\sqrt{\pi n}$ ) held constant during this “charging process”. As  $\lambda$  increases to unity, the PDFs  $g_{ij}(\lambda, r)$ , where  $i, j$  indicate spin species, change from the non-interacting  $g_{ij}^0(r)$ , Fig. 2, to the interacting from  $g_{ij}(r)$ . This is accompanied by the “build up” of the energies arising from the interactions. Thus the exchange-correlation energy  $E_{xc}$  appears as an integration over the coupling constant (if we are dealing with a system at a finite temperature  $T$ , then we have Helmholtz free energies  $F_{xc}$  instead of  $E_{xc}$ ). Thus, if  $x_1, x_2$  are the fractional compositions  $n_i/n$ , then the exchange-correlation energy per particle  $E_{xc}/N$  is

$$E_{xc}/N = \int_0^1 \frac{d\lambda}{\lambda} \frac{n}{2} \int 2\pi r dr \frac{\lambda}{r} \sum_{ij} x_i x_j (g_{ij}(\lambda, r) - 1). \quad (1)$$

This is known as the adiabatic connection formula [11], and is due to Pauli. Thus if the fully interacting PDF is known, the exchange-correlation energy can be obtained without the limitations of perturbation theory. The full PDF can be evaluated by (i) Quantum Monte Carlo methods (QMC), (ii) integral-equation methods developed from correlated wavefunctions using Feenberg-type analyses, (iii) classical-map hyper-netted chain (CHNC) methods, (iv) methods for specific applications as in Giorgi et al. [10] (v) SLTS methods for moderately coupled fluids.

### The Exchange Energy

Note that the exchange energy alone is given by

$$E_x/N = \frac{n}{2} \int 2\pi r dr v(r) \sum_{ij} x_i x_j (g_{ij}^0(r) - 1). \quad (2)$$

That is, the Fermi hole of the non-interacting PDF determines the exchange energy. This was *defined* as a contribution taken to *first order*, Fig. 3b. Since one factor of  $v(r)$  is already included in Eq. (2), we can only use  $g_{ij}^0(r)$  as a better form would have gone beyond first order in  $V_{ee}$ . The noninteracting PDF  $g_{ij}^0(r)$  can be evaluated [18] directly from the noninteracting one-body wavefunctions (plane

waves):

$$\begin{aligned} h_{ij} &= g_{ij} - 1 \\ h_{ij}^0(r) &= -\delta_{ij} F_i(r) F_j(r) \\ F_i(r) &= \frac{1}{n_i} \int \frac{d\vec{k}}{(2\pi)^2} f_i(k) e^{i\vec{k}\cdot\vec{r}}. \end{aligned}$$

Here  $f_i(k)$  is the Fermi occupation number for the momentum state  $k$ . At  $T = 0$ ,  $F_i(r) = 2J_1(k_i r)/(k_i r)$ ;  $k_i = \sqrt{4\pi n_i}$  where  $J_1(x)$  is a Bessel function [1], and  $k_i$  is the Fermi momentum of the spin species  $i$ . At finite temperatures, the integrations have to be done numerically.

### Formulae for the Kinetic and Exchange Energies.

At zero temperature, the Hartree–Fock energy can be written in terms of the spin polarization  $\zeta$  and the density parameter  $r_s$ . The Hartree energy  $E_H$  is zero.

$$E_{HF} = E_0 + E_H + E_X; \quad E_H = 0, E_0/N = \frac{1 + \zeta}{2r_s^2};$$

$$E_x/N = -\frac{2\sqrt{2}}{3\pi r_s} [(1 + \zeta)^{3/2} + (1 - \zeta)^{3/2}]$$

The spin-dependent energies, e. g.  $E_x(\zeta)$ , may be written as:

$$\begin{aligned} E_x(\zeta) &= E_x(0) + (E_x(1) - E_x(0))P(\zeta) \\ P(\zeta) &= \frac{\zeta_+^\alpha + \zeta_-^\alpha - 2}{2^\alpha - 2}. \end{aligned}$$

Here  $\zeta_\pm = (1 \pm \zeta)$ , and  $P(\zeta)$  is called the “polarization function”;  $\alpha = 3/2$  in Hartree–Fock theory.

### The Correlation Energy

The calculation of the correlation energy  $E_c$  is the heart of the many-body problem, and is the main challenge. If  $E_c$  were known, the total ground state energy is  $E_T = E_{HF} + E_c$ . A widely used QMC evaluation was given in 1989 by Tanatar and Ceperley [23] using a variational Monte Carlo method (VMC) as well as a Greens Function Monte Carlo method. They gave parametrized forms for the unpolarized ( $\zeta = 0$ ) and fully polarized ( $\zeta = 1$ ) energies per electron, i. e.,  $\epsilon_c(r_s)$ , in the form

$$\epsilon_c(r_s) = a_0 \frac{1 + a_1 x}{1 + a_1 x + a_2 x^2 + a_3 x^3}$$

where  $x = \sqrt{r_s}$ . The parameters have the values,  $a_0 \dots a_3 = -0.1784, 1.1300, 0.9052, 0.4165$  for  $\zeta = 0$  and  $-0.02575, 340.5813, 75.2293, 37.0170$  for  $\zeta = 1$ . No simulations were done for intermediate  $\zeta$ , and hence it

was often *assumed* that the exchange-like polarization function  $P(\zeta)$  could be used for the correlation energy as well.

$$\epsilon_c(\zeta) = \epsilon_c(0) + (\epsilon_c(1) - \epsilon_c(0))P(\zeta)$$

A more detailed study, given by Attacalite et al. in 2002 is currently the best available QMC form of  $E_c$ . These authors [3] did simulations for finite- $\zeta$  as well and proposed a parameterized form.

$$\begin{aligned}\epsilon_c(r_s, \zeta) &= (e^{-\beta r_s} - 1)\epsilon_x^6(r_s, \zeta) \\ &= +\epsilon_c(r_s, 0) + \alpha_1(r_s)\zeta^2 + \alpha_2\zeta^4 \\ \alpha_1(r_s) &= (1/2) \left[ \partial^2 \epsilon_c(r_s, \zeta) / \partial^2 \zeta^2 \right]_{\zeta=0} \\ \alpha_2(r_s) &= (1/24) \left[ \partial^4 \zeta^4 \right]_{\zeta=0} \\ \epsilon_x^6(r_s, \zeta) &= \epsilon_x(r_s, \zeta) - \{1 + (3\zeta^2)(1 + \zeta^2/16)\}\end{aligned}$$

Here parameters  $\alpha_1, \alpha_2$  are spin-stiffness constants. A short computer program for the calculation of  $\epsilon_c$  is available [24].

### CHNC Approaches

QMC procedures are computationally very demanding, and hence the development of other, simpler methods is of great interest. The accuracy of such methods can be ascertained by comparison with QMC results, and then such methods can be used in areas where QMC is prohibitive (e.g., at finite temperatures, many valley systems). In this respect, the classical-map HNC approach (CHNC) is worthy of note. In CHNC [21], the quantum fluid at  $T = 0$  is replaced by a classical Coulomb fluid at a finite temperature  $T_q$ . The latter is chosen requiring that (i) the non-interacting classical fluid has the same  $g_{ii}^0(r)$  as the quantum fluid. (ii) the correlation energy of the interacting classical fluid has the same correlation energy as the quantum fluid. The first requirement is easily met, and involves the construction of a classical potential which would give the same Fermi hole as the quantum fluid (see Fig. 2). The classical potential,  $\beta V_{\text{pau}}(r)$  is called the “Pauli potential” and is a universal function of  $rk_F$ . This potential is zero for anti-parallel spins. This procedure fixes a ‘Pauli potential’ but not the temperature  $T_q$ , since only the product  $\beta \times V_{\text{pau}}(r)$  is obtained. The second condition is used to determine  $T_q$  by evaluating  $\epsilon_c(r_s)$  using the adiabatic connection formula, for a trial value of  $T_q$ , and a chosen  $\zeta$ , say  $\zeta = 1$ . Then  $T_q$  is adjusted till the QMC value of  $\epsilon_c(r_s, \zeta = 1)$  is obtained. Thus a table of  $T_q$  for each  $r_s$ , i.e.,  $T_q(r_s)$ , is obtained. The classical PDFs needed here are easily calculated using a well established hyper-netted-chain technique. Only the  $\epsilon_c$  value at  $\zeta = 1$  is used as the

input to the fit. The output is the full tabulation of the interacting  $g_{ij}(r)$ . It is found to be in *remarkably good agreement with the QMC generated  $g_{ij}(r)$* . This is perhaps a consequence of DFT where it is asserted that the true density distribution is obtained if the true ground state energy (i.e.,  $E_0 + E_X + E_c$ ) is captured. It is also found that this CHNC procedure yields accurate  $\epsilon_c(r_s)$  values at spin-polarizations different to the input value. The  $T_q(r_s)$  function is found to be transferable to many Coulomb-fluid problems including hydrogen plasmas [6].

### Fermi-Liquid Parameters

A knowledge of the correlation energy as a function of the  $r_s, \zeta$ , and  $T$  enables an easy evaluation of Landau Fermi liquid parameters (LFLP). These are usually evaluated from complicated perturbation theory calculations whose domain of validity is, strictly speaking, restricted to small  $r_s$ . Three quantities are of great interest for the LFLP. These are, the inverse compressibility  $\kappa$  which is the density derivative of the chemical potential  $\mu$ , the enhanced spin-susceptibility  $\chi$  compared to the Pauli susceptibility  $\chi_P$ , (incorporated in  $g^*$ , the effective Landé  $g$  factor), and the effective mass  $m^*$ . The latter involves the interacting and ideal specific heats  $C_v, C_v^0$ . At finite  $T$ , the correlation energy is replaced by the correlation Free energy,  $F_c(r_s, \zeta, T)$ . This is the correction to the Helmholtz free energy beyond the Hartree-Fock approximation. If  $F_c(r_s, \zeta, T)$  is known, a non-perturbative result becomes available for  $\kappa, m^*$  and  $g^*$  (for detailed, see [8]). Thus:

$$\begin{aligned}m^* &= C_v/C_v^0 = 1 + \frac{[\partial^2 F_{xc}(t)/\partial t^2]}{[\partial^2 F_0(t)/\partial t^2]} \\ \chi_P/\chi_s &= (m^* g^*)^{-1} = 1 + \frac{[\partial^2 F_{xc}(\zeta)/\partial \zeta^2]}{[\partial^2 F_0(\zeta)/\partial \zeta^2]} \\ 1/\kappa &= n^2 \partial \mu / \partial n = n^2 \partial^2 [F_0 + F_X + F_c] / \partial n^2.\end{aligned}$$

Calculations of the ground-state energy as a function of  $r_s, \zeta$ , as well as the spin-susceptibility enhancement using the QMC and the CHNC methods show that the 2D fluid (which is paramagnetic  $\zeta = 0$  at high densities), becomes a ferromagnetic liquid (see [3,7]) for  $r_s \simeq 26$ . Thus the gain in kinetic energy due to spin-polarization is compensated by the exchange-correlation energy at sufficiently low densities. If simple perturbation methods are used for evaluating  $E_{xc}(r_s)$ , incorrect predictions (e.g., a magnetic transition at  $r_s \sim 2$  in RPA) are obtained. The spin-susceptibility enhancement “blows up” towards infinity, as the transition is approached. These theoretical predictions have been confirmed experimentally (see [22]).

## 2D Layers with Finite Thickness

The assumption that the 2D layer is infinitely thin (i. e., “ideal”) is often incorrect in practice. The lowest subband of a quantum well supporting a 2D layer may have an  $z$ -extension  $w$  of 10–20 nm, with a density distribution  $n(z)$ . If the next subband is inaccessible to the electrons, then they are 2D systems with a finite thickness. The HIGFETS used in many experiments contain a nearly triangular potential well (see Fig. 1) and the lowest subband wavefunction is approximated by the Fang-Howard(FH) function [2]  $\phi_{\text{fh}}(z)$ . The difficulty of modeling exchange and correlation in such systems using either the ideal-2D parametrized  $E_{\text{xc}}$  forms, or the 3D forms (to allow for the  $z$ -extension) was noted by Martin et al. [14]. In fact, new exchange-correlational functionals have to be constructed for the effective-2D potential found in thick-2D layers [5]. We denote the Coulomb potential in an infinitely thin layer by  $V(r) = 1/r$ , while  $W(r)$  is used for the effective 2-D potential of a thick layer. The potential  $W(r)$  between two electrons having coordinates  $(\vec{r}_1, z_1)$  and  $(\vec{r}_2, z_2)$ , with  $\vec{r} = \vec{r}_1 - \vec{r}_2$  is given by,

$$W(r) = \int_0^{z_m} \int_0^{z_m} \frac{dz_1 dz_2 n(z_1) n(z_2)}{[r^2 + (z_1 - z_2)^2]^{1/2}}. \quad (3)$$

Here  $z_m$  is  $\infty$  for FH, while  $z_m = w$  for a quantum well. The potential  $W(r) = (1/r)F(r)$  and the form factor  $F(r)$  reflects the effect of the  $z$ -extension of the density. It has been shown that any arbitrary density distribution  $n(z)$  can be replaced by a electrostatically equivalent rectangular slab of width  $w$  by choosing  $w$  such that: in the 2D plane as  $n(z)$ .

$$n_{\text{cd}} = 1/w = \int n(z)^2 dz. \quad (4)$$

The quasi-2D potential for a constant-density slab of width  $w$  is given by

$$W(r) = V(r)F(s), \quad s = r/w, \quad t = \sqrt{1 + s^2} \quad (5)$$

$$F(s) = 2s \left[ \log \frac{1-t}{s} + 1 - t \right]. \quad (6)$$

This potential tends to  $1/r$  for large  $r$ , and behaves as

$$\frac{2}{w} \left( \ln \frac{2w}{r} + \frac{r}{w} - 1 \right)$$

for  $r < w$ . Thus the short-range behaviour is logarithmic and weaker than the Coulomb potential. The  $k$ -space form

of the CDM potential is:

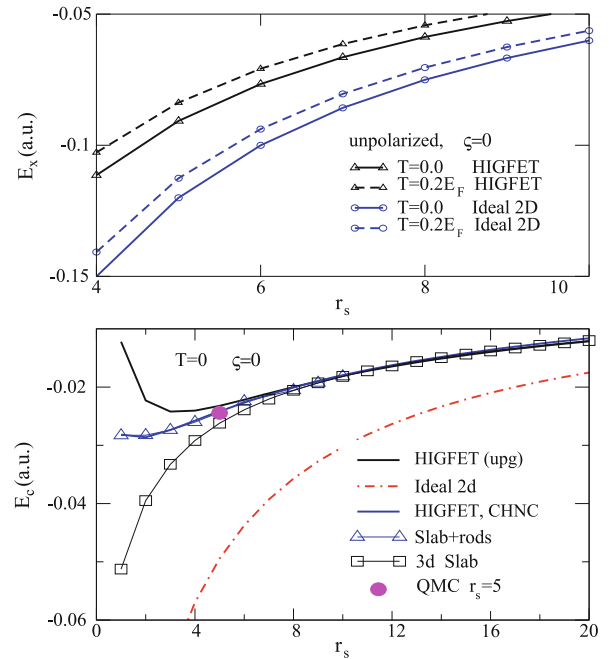
$$W(k, w) = V(k)F(p), \quad p = kw \quad (7)$$

$$F(p) = (2/p) \{ (e^{-p} - 1)/p + 1 \}. \quad (8)$$

The form factors  $F(s)$  and  $F(p)$  tend to unity as  $w \rightarrow 0$ . The effective  $w$  in HIGFETS depends on  $r_s$ , i. e., in atomic units.

$$w = 16/(3b)b^3 = 33/(2r_s^2).$$

Using these potentials, layer-thickness dependent exchange-correlations functionals  $E_{\text{xc}}(r_s, \zeta, w)$  can be evaluated using perturbation methods, or using the CHNC. The resulting xc-functionals have been given in [5], and displayed in Fig. 4. It is found that the exchange energy is reduced due to layer thickness which enters as  $w/r_s$ , thus



**Spin Dependent Exchange and Correlation in Two-Dimensional Electron Layers, Figure 4**

**a** The exchange energy  $E_x$  (Hartree a.u.) of a 2DES in a HIGFET compared to that of an ideal 2DES, at  $T/E_F=0$  and 0.2, with  $\zeta = 0$ . Solid line with circles, ideal 2D,  $T = 0$ , Solid line with triangles, HIGFET at  $T = 0$ . Corresponding broken lines are for  $T = 0.2E_F$ . **b** The correlation energy  $E_c$  at  $T = 0$  and  $\zeta = 0$ , for the HIGFET layer. HIGFET(upg), black solid line, is the “unperturbed-g” approximation. The deep grey dashed line, HIGFET(CHNC), is the full calculation. This is compared with the correlation energy of a 3D slab model (line with squares), and the “slab+rod” model (triangles). For details, see [5]. The QMC datum (black hatched circle) for a HIGFET,  $r_s = 5$ , is from [22]

affecting high densities and thick layers. A most interesting observation is that the transition to a ferromagnetic state which occurs at  $r_s \simeq 26 - 27$  in ideal 2D, is pushed to higher  $r_s$  as the thickness  $w$  is increased.

### Multi-Component 2D Layers

The two most important multi-component systems are 2D layers found in Si/SiO<sub>2</sub> interfaces, and in graphene monolayers.

#### Two-Valley System in Silicon-Silica Interfaces

Bulk Si has six equivalent valleys. However, when SiO<sub>2</sub> layers are grown on the (001) surface, with  $z$  the growth direction, the 2D layer occupies two degenerate layers in the (001) plane. This is a 4-component fluid, with two spin states and two valley states. Thus using an index  $\nu = 1, 2, 3, 4$  where the first two are for spin, and the last two are for the valleys, there are 10 PDFs in the upper triangle of the matrix  $g_{\nu\eta}(r)$ . Only five of these are independent even for  $\zeta \neq 0$ . Four of these 10 are diagonal and contribute to both exchange and correlation, where as the off diagonals contribute *only* to the correlation energy  $E_c$ . This contrasts with the usual single valley system where there are two diagonals, and one off-diagonal PDF in the matrix of PDFs. Thus correlation effects dominate over exchange effects in multi-valley systems, and they show less tendency to spin-polarization effects which are driven by exchange effects. Thus there is no spontaneous spin-polarized state in the 2D-two valley system. However, the large increase in the correlation energy has been argued to be responsible for a rapid increase in the effective mass of electrons in Si MOSFETS as the density  $n$  decreases towards  $\simeq 8 \times 10^{10}$  electrons/cm<sup>2</sup> [8,15]. However, this is found to be sensitive to slight changes in the CHNC parametrization.

Quantum Monte Carlo results of  $E_{xc}(r_s)$  at  $\zeta = 0, 1$  and CHNC results (at many values of  $\zeta$ ) are available, and are in excellent agreement with each other. If the total electron number density per unit area in the 2D layer is  $n$  in atomic units, then the overall  $r_s = 1/\sqrt{\pi n}$ , while the valley density,  $n/2$  yields an  $r_{sv} = \sqrt{2}r_s$ . This  $r_{sv}$  and the corresponding  $k_{FV}$  are used in constructing the non-interacting PDFs  $g_{\nu\eta}^0$  from which the exchange energy can be calculate. Clearly, this is  $2\epsilon_x(r_{sv})$ , where  $\epsilon_x$  is the usual one-valley formula for  $E_x/N$ .

The correlation energy cannot be completely mapped into results based on the 1-valley system. Thus  $E_c$  is determined via a coupling constant integration over the 10 PDFs, obtained from QMC or CHNC, using the adiabatic

connection formula. Detailed results and numerical fits are available in [8].

### Graphene: 2D Two-Valley System on a Honey-Comb Lattice

The electrons in the two-valley 2D system of Si/SiO<sub>2</sub> interfaces carry a valley index, but the single-particle wavefunctions are simple plane waves. In graphene, the two-component wavefunctions of massless fermions have a “chiral” character, since they carry phase angles  $\phi_k$ , where  $\vec{k}$  is the momentum vector in the 2D plane. The valence and conduction bands touch at the  $\mathbf{K}, \mathbf{K}'$  points, and we need a band index  $b = 1, -1$ , to indicate the conduction and valence bands. Thus, we may have simultaneous electron and hole occupations, leading to an eight-component problem. Although the  $\epsilon_x$  can be calculated explicitly, the calculation of the  $\epsilon_c$  becomes very demanding. However, as the coupling constant  $g$  is independent of the electron density, being  $\simeq 2.7/\epsilon$ , the interactions are weak. Also, for pure electron doping or pure hole doping, the  $\epsilon_c$  of the 4-component system in Si/SiO<sub>2</sub> at the same coupling strength as  $g$  (i.e,  $r_s = g$ ) may be used as an estimate. Hence we focus on  $E_x$  and introduce a unit of energy  $E_u = v_F k_c$ . Here  $v_F$  is the Fermi velocity of the massless fermions, and  $k_c$  is a cutoff wavenumber chosen so that the number of states in the Brillouin zone is conserved, that is, if  $A_0$  is the area per carbon atom,  $k_c^2 = 4\pi(1/A_0)$ . Then the exchange energy per Carbon atom can be written as:

$$E_x/E_u = -\frac{A_0 g^0/k_c}{(2\pi)^2} \frac{1}{4} \sum_{b_1, b_2, \sigma} \int_0^{2\pi} d\theta dk dp \times kp \frac{1 + b_1 b_2 \cos(\theta)}{|\mathbf{k} - \mathbf{p}|} n_{b_1, \sigma}(k) n_{b_2, \sigma}(p). \quad (9)$$

In the above  $g^0 = 2.7/\epsilon$ . A special feature, not found in the usual (non-chiral) 2D system is the occurrence of a term which depends on  $\theta$ , the scattering angle. Because of this modulation of the interaction potential, the exchange energy in graphene is weaker than in a normal 2D layer with the same coupling strength. The above expressions for the exchange energy may be expressed [19] in terms of integrals over PDFs involving Bessel and Struve functions, or as elliptic elliptic integrals which depend on  $\zeta$  and the fractional doping. Calculations of the total energy, (i.e, kinetic+ $E_{xc}$ ) for graphene layers doped with electrons, holes or both, *do not* support a transition to a spin-polarized ground state. This is not surprising since the strength of the coulomb interaction is  $\sim 2.7/\epsilon$ , i.e., weak, given that the spin transition in usual 2D systems occur for  $r_s \sim 26$ .



### Exchange and Correlation in a Magnetic Field

Electrons respond to an external magnetic field by forming Landau Levels. If  $B$  is the magnetic field in the  $z$ -direction, then the energy spectrum is that of a harmonic oscillator, where the cyclotron frequency  $\omega_c$  and the magnetic length  $\ell$  are given by:

$$\omega_c = |e|B/(\hbar m_b), \quad \ell = \{\hbar c/|e|B\}^{1/2}. \quad (10)$$

Here  $m_b$  is the band mass and  $c$  is the velocity of light. The Harmonic oscillator energy levels with index  $\eta$  are:

$$\epsilon_\eta = (\eta + 1/2)\hbar\omega_c + \epsilon_{sb} \quad (11)$$

contains the lowest subband energy  $\epsilon_{sb}$  of  $z$ -motion, and this may be taken as the zero of energy. The momentum states of the 2D system collapse into degenerate states, with each Landau level having  $N_L = eBA/\hbar c$  states, where  $A$  is the area of the system. If there are  $N$  electrons in  $A$ , the filling factor  $\nu = N/N_L = n/(2\pi\ell^2)$ . The integer part  $\eta$  of  $\nu$  gives the number of filled Landau levels, while  $\nu - \eta$  is the part of the highest Landau level which is partially filled. The filled Landau levels are like "inert shells" of atoms, while the partially filled part is polarizable and accounts for the more interesting aspects of exchange and correlation among electrons in magnetic fields. The total filling factor  $\nu = \nu_1 + \nu_2$  where  $\nu_i$  is for the  $i$ th spin species.

Since we are interested in  $\epsilon_{xc}$ , we need the PDFs for use in Eq. (1). Thus it is convenient to work in the symmetric gauge where the eigenstates are functions of the radial distance  $r$  and the angle  $\phi$  in the 2D plane, having radial and angular momentum quantum numbers  $n_r$  and  $m$ . Then, with  $\hbar = 1$ ,

$$\epsilon_{n_r, m} = \{n_r + (|m| - m)/2 + 1/2\}\omega_c. \quad (12)$$

When the field is weak, there are a large number of Landau levels, and the usual 2D formulae for exchange and correlation can be used. This can be improved by using the summations over the density of states instead of integrations over  $k_x, k_y$ . A discussion may be found in, e.g., [4]. The high magnetic field limit, where only a few occupied Landau levels is of great importance. The integer and fractional QHE are found in this 'high-field' regime. If the number of full Landau levels of spin  $\sigma$  is  $\nu_\sigma$ , the non-interacting PDF  $g^0(r)$  is:

$$g^0(r) = 1 - \frac{e^{-x/2}}{\nu^2} \sum_{\sigma} [L_{\nu_\sigma-1}^1(x^2/2)]^2 \quad (13)$$

where  $x = r/\ell$ , and  $L_n^1$  is an associated Laguerre polynomial. The spin summation is unimportant in high fields,

when the LLL is fully polarized by the field.  $\epsilon_x$  can be evaluated from Eq. (2). The 2D Coulomb interaction may contain finite-layer thickness corrections. The effect of the magnetic field is totally contained in  $e_{\text{mag}} = e^2/(\epsilon\ell)$  and forms a natural energy unit for this system. The exchange energy  $\epsilon_v$  (in  $e_{\text{mag}}$  units), is  $-\sqrt{\pi/8}$  at  $\nu = 1$ , and increases almost linearly to  $2.35\epsilon_1$  at  $\nu = 6$ .

The correlation energy  $\epsilon_c(r_s, \nu)$  of filled Landau levels can be evaluated using the adiabatic connection formula and the interacting PDFs. However, only an RPA evaluation is available, and  $\epsilon_c(r_s, \nu)$  is very close to the zero. Hence the  $B = 0$  value of  $\epsilon_c$  is a good approximation.

The highest occupied Landau level is partially filled if  $\nu_L = \nu - \eta$  is nonzero. The correlations in partially filled states are clearly manifest in the high field limit when the lowest Landau level is fractionally occupied. Here  $\epsilon_c$  has to be calculated from Quantum Monte Carlo simulations, or from methods based on the plasma analogy, via an ansatz for the many particle wavefunction. The case  $\nu_L = 1/3$  and other "odd-integer" fractional-QHE states were elegantly explained by Laughlin [4], leading to a Nobel prize. This can be extended to other fractions within the plasma analogy to construct PDFs and obtain  $\epsilon_c$  via the adiabatic connection formula [17]. Quantum simulations as well as composite-Fermion models have now largely superseded such methods. The literature on the quantum Hall effect [4] should be consulted for more details.

### Exchange and Correlation at Finite Temperatures

The first-order (exchange) free energy  $F_x$  consists of  $F_x^i$ , where  $i$  denotes the two spin species. At  $T = 0$  these reduce to the exchange energies:

$$E_x^i/n = \frac{8}{3\sqrt{\pi}} n_i^{1/2}. \quad (14)$$

Here  $n_1 = n(1 + \zeta)/2$ , and  $n_2 = n(1 - \zeta)/2$ . Then the exchange energy per particle at  $T = 0$ , i.e., the total  $E_x/n$  becomes

$$E_x/n = (E_1^x + E_2^x)/n = -\frac{8}{3\pi r_s} [x_1^{3/2} + x_2^{3/2}] \quad (15)$$

where  $x_1$  and  $x_2$  are the fractional compositions  $(1 \pm \zeta)/2$  of the two spin species.

We define a reduced temperature  $t = T/E_F$ ,  $E_F = \pi n$ , and the species-dependent reduced chemical potentials  $\mu_i^0/T$  by  $\eta_i$ , reduced temperatures  $t_1 = t/(1 + \zeta)$  and  $t_2 = t/(1 - \zeta)$ , based on the two Fermi energies  $E_{F1}$  and  $E_{F2}$  which are  $E_F(1 \pm \zeta)$ . Then we have:

$$F_x^i/E_x^i = \frac{3}{16} t_i^{3/2} \int_{-\infty}^{\eta_i} \frac{I_{-1/2}^2(u) du}{(\eta_i - u)^{1/2}}. \quad (16)$$

The  $I_{-1/2}$  is the Fermi integral defined as usual:

$$I_\nu(z) = \int_0^\infty \frac{dx x^\nu}{1 + e^{x-z}}. \quad (17)$$

The  $\eta_i$  are given by

$$\eta_i = \log(e^{1/t_i} - 1). \quad (18)$$

In the paramagnetic case Eq. (16) reduces to the result given by Isihara et al. [13]. For small values of  $t$ , the exchange energy is of the form,

$$E_x(r_s, t) = E_x(r_s, 0)[1 + (\pi^2/16)t^2 \log(t) - 0.56736t^2 + \dots]. \quad (19)$$

The total exchange free energy is  $F_x = \Sigma F_i^x$ .

A real-space formulation of  $F_x = F_1^x + F_2^x$  using the zeroth-order PDFs fits naturally with the approach of our study. Thus

$$F_x/n = n \int \frac{2\pi r dr}{r} \sum_{i<j} h_{ij}^0(r). \quad (20)$$

Here  $h_{ij}^0(r) = g_{ij}^0(r) - 1$ . In the non-interacting system at temperature  $T$ , the antiparallel  $h_{12}^0$ , viz.,  $g_{12}^0(r, T) - 1$ , is zero while

$$h_{11}^0(\mathbf{r}) = -\frac{1}{n_i^2} \Sigma_{\mathbf{k}_1, \mathbf{k}_2} n(k_1)n(k_2)e^{i(\mathbf{k}_1 - \mathbf{k}_2) \cdot \mathbf{r}} = -[f(r)]^2.$$

Here  $\mathbf{k}$ ,  $\mathbf{r}$  are 2D vectors and  $n(k)$  is the Fermi occupation number at the temperature  $T$ . At  $T = 0$   $f(r) = 2J_1(k_i r)/k_r$  where  $J_1(x)$  is a Bessel function.

The exchange free energy is a universal function  $F_x(t)/E_x$ , for arbitrary  $\zeta$ . That is, the same function applies to any component, on using the reduced Fermi temperature of the spin species. The total  $F_x$  is the sum of both spin terms. A parametrized fit is:

$$F_i^x(t, \zeta)/E_i^x(\zeta) = \frac{1 + C_1 t_i + C_2 t_i^2}{1 + C_3 t_i + C_4 t_i^2} \tanh(1/\sqrt{t_i}). \quad (21)$$

The fit coefficients  $C_i$  are 3.27603, 4.81484, 3.33100, 6.51436. The temperature  $t_i$  is  $t/(1 \pm \zeta)$ , appropriate to the spin polarization. The exchange effects in the 2DES decay more slowly with temperature than in the 3D case where a  $\tanh(1/t)$  factor appears in Eq. 3.2 of [20]. The above form does not explicitly contain the low-temperature logarithmic term [12], but it reproduces the value of

0.99382 at  $t = 0.05$ , while the numerical integration gives 0.9939497. Similarly, at  $t = 1, 10$  and  $30$  the fit (integral) returns 0.63839 (0.63839), 0.22999 (0.22990), and 0.13421 (0.13410) respectively.

If the 2D layer has a finite thickness,  $E_x$  is reduced. For numerical procedures and parametrized forms, the reader is referred to [5]. The correlation energy  $\epsilon_c(r_s, \zeta, T)$  can be easily calculated using the PDFs from CHNC [7], but they have not been presented in a parametrized form.

## Conclusion

An understanding of electron-exchange and correlation, i.e., many-body effects, is crucial to the design of quantum-well lasers, field-effect transistors, spintronics and other areas of nanostructure technology. We currently have excellent results for uniform electron systems at zero magnetic fields. The case of moderate magnetic fields is poorly understood. The very high-field case has progressed with our understanding of the fractional and integer quantum Hall effects.

## Bibliography

1. Abramovitz M, Stegun IA (1965) Handbook of mathematical functions. Dover, New York
2. Ando T, Fowler B, Stern F (1982) Rev Mod Phys 54:437
3. Attacalite C, Moroni S, Gori-Giorgi P, Bachelet GB (2002) Phys Rev Lett 88:256601; (2003) Erratum 91:109902
4. Chakraborty T (1996) The fractional quantum hall effect. Springer, New York
5. Dharma-wardana MWC (2005) Phys Rev B 72:125339
6. Dharma-wardana MWC, Perrot F (2002) Phys Rev B 66:14110
7. Dharma-wardana MWC, Perrot F (2003) Phys Rev Lett 90:136601
8. Dharma-wardana MWC, Perrot F (2004) Phys Rev B 70:035308
9. Geim AK, Novoselov KS (2007) Nat Mater 6:183–191
10. Gori-Giorgi P, Perdew JP (2001) Phys Rev B 64:155102
11. Harris J, Jones RO (1974) J Phys F 4:1170
12. Hong S, Mahan GD (1995) Phys Rev B 52:7860
13. Isihara A, Toyoda T (1980) Phys Rev B 21:p3358
14. Kim Y-H, Lee I-H, Nagaraja S, Leburton J-P, Hood RQ, Martin RM (2000) Phys Rev B 61:5202
15. Kravchenko SV, Sarachik MP (2004) Rep Prog Phys 67:1
16. Liu HC, Song CY, Wasilewski ZR, Springthorpe AJ, Cao JC, Dharma-wardana C, Aers GC, Lockwood DJ, Gupta JA (2003) Phys Rev Lett 90:077402
17. MacDonald AH, Aers GC, Dharma-wardana MWC (1985) Phys Rev B 31:5529
18. Mahan GD (1981) Many-particle physics. Plenum, New York
19. Peres NMR, Guinea F, Castro Neto AH (2007) Phys Rev 72:174406; Dharma-wardana MWC (2007) Ibid 75:075427
20. Perrot F, Dharma-wardana MWC (1984) Phys Rev A 30:2619
21. Perrot F, Dharma-wardana MWC (2001) Phys Rev Lett 87:206404
22. Tan YW, Zhu J, Stormer HL, Pfeiffer LN, Baldwin KN, West KW (2005) cond-matt/0412260, Phys Rev Lett 94:16405; De Palo S,

Botti M, Moroni S, Senatore G (2005) Phys Rev Lett 94:226405;  
 Vakili K, Shkolnikov YP, Tutuc E, DePoortere EP, Shayegan M  
 (2004) Phys Rev Lett 92:226401

23. Tanatar B, Ceperley DM (1989) Phys Rev B 39:5005  
 24. chandre.dharma-wardana@nrc-cnrc.gc.ca, <http://babylon.phy.nrc.ca/ims/qp/chandre/>

## Spin Dynamics in Disordered Solids

FRIDRIKH S. DZHEPAROV

Institute for Theoretical and Experimental Physics,  
 Moscow, Russia

### Article Outline

Glossary

Definition of the Subject

Introduction

Delocalization of Nuclear Polarization  
 in a Disordered Spin System

Nuclear Relaxation via Paramagnetic Impurities

Resonance Line Form Function

for Magnetically Diluted Solids

Saturation on the Wing of a Dipole-Broadened EPR Line  
 and Cluster Expansions

Future Developments

Bibliography

### Glossary

**Averaging** Two kinds of averaging are necessary for any observable  $\hat{Q}$ : a standard quantum mechanical one  $Q = \langle \hat{Q} \rangle = \text{Tr}(\hat{Q}\rho)$ , where  $\rho$  is the density matrix, and consequent averaging over all possible positions of particles in disordered media  $\langle Q \rangle_c = \langle \langle \hat{Q} \rangle \rangle_c$ .

**Continuum media approximation** Positions of particles forming the disordered solids can be considered as a subset (impurity sites) of crystal lattice sites, randomly distributed on the lattice with a small probability  $c \ll 1$  to find a given site occupied. Many important results can be received in continuum media approximation (CMA) when prime cell volume  $\Omega_c \rightarrow 0$  together with impurity concentration  $c \rightarrow 0$  at a fixed value of impurity density  $n = c/\Omega_c$ .

**Disordered solids** Statically disordered media are considered, this means that constituent particles (atoms, ions and so on) do not participate in significant translational motions during the relaxation time under discussion, and therefore their positions are fixed (frozen) in the main approximation.

**High-temperature approximation** High-temperature approximation (HTA) in spin dynamics consists of using the simplest density matrix  $\rho = 1/\text{Tr} 1$ , corresponding to the limit of infinite temperature  $T$  of canonical Gibbs distribution  $\rho_G = \exp(-H/T)/\text{Tr} \exp(-H/T)$ ; here  $H$  is the Hamiltonian. Spin dynamics remains nontrivial and rich in this limit. Sometimes HTA means application of  $\rho = (1 - H/T)/\text{Tr} 1$ .

**Local field** See “Secular part of dipole–dipole interactions”.

**Local frequency** The frequency  $\omega_{\text{loc}} = \gamma H_{\text{loc}}$  of rotation of the spin in local field  $H_{\text{loc}}$ .

**Secular part of dipole–dipole interactions** As a rule any spin is considered as subjected to an external static magnetic field  $\mathbf{H}_0 = H_0 \mathbf{n}_z$  (directed along the  $z$ -axis) and local field, produced by dipole magnetic moments of surrounding spins. The Hamiltonian of the dipole–dipole interaction of spins  $\mathbf{I}_i$  and  $\mathbf{I}_j$  is of the form  $H_d^{(0)}(i, j) = r_{ij}^{-3}((\mathbf{m}_i \mathbf{m}_j) - 3(\mathbf{n}_{ij} \mathbf{m}_i)(\mathbf{n}_{ij} \mathbf{m}_j))$ , where  $\mathbf{r}_{ij} = \mathbf{n}_{ij} r_{ij} = \mathbf{r}_i - \mathbf{r}_j$ , and  $\mathbf{r}_i$  is the position of spin “ $i$ ” having magnetic moment  $\mathbf{m}_i = \hbar \gamma_i \mathbf{I}_i$ . If  $H_0 \gg H_{\text{loc}}$ , where  $H_{\text{loc}}$  is the mean square field produced at any spin by surrounding spins (local field), then the Hamiltonian  $H_d^{(0)}(i, j)$  can be substituted by the so-called secular dipole–dipole Hamiltonian  $H_d(i, j) = \frac{1}{2r_{ij}^3}(1 - 3(\mathbf{n}_{ij} \mathbf{n}_z)^2)(\mathbf{m}_i \mathbf{m}_j - 3(\mathbf{m}_i \mathbf{n}_z)(\mathbf{m}_j \mathbf{n}_z))$ . The accuracy of the substitution is not less than  $\sim H_{\text{loc}}/H_0$ .

**Spin dynamics** Spin dynamics is considered as a time evolution of correlation functions directly connected with measurable quantities of paramagnetic samples.

**Units** As a rule  $\hbar = 1$  is supposed. Part of the equations contains  $\hbar$  written explicitly.

### Definition of the Subject

The main specifics of the theory of spin dynamics in disordered solids result from the fact that calculation of observable values must start from the solution of the equation of motion, and then they should be averaged over random distribution of spins in the sample. Nominally any problem in statistical physics looks analogous, but the content of existing text books contains much more simple bypass methods to achieve the results. An important bypass class, determined in the preceding century, consists of problems that are equivalent to the motion of weak (or seldom) interacting (quasi)particles in translational invariant media. The basis of this solution is formed by the Boltzmann equation (devised in the 19th century) and by methods of deriving hydrodynamic equations based on this foundation [1]. Other important advancements in physical kinet-

ics are connected with the invention of a projection technique by Nakajima–Zwanzig for deriving various master equations and with the development of effective approximations for corresponding memory functions [1,2]. Both these bypass methods produce satisfactory predictions if the system has such small parameters as, for example, the ratio of collision duration to time between collisions, or ratio of memory time to time of variation of substantial observables (the observables whereon the system evolution is projected). Similar small parameters do not exist in disordered solids after a beginning stage, or they are totally absent. Therefore, other more refined methods are necessary to predict experimental results. They are partially presented below.

## Introduction

The first correct solutions of the problems of disordered media kinetics, which are directly connected with the spin dynamics, were obtained by Forster [3] and Anderson [4].

Forster's problem (in application to spin kinetics) describes the relaxation of impurity nuclei via paramagnetic impurities in the absence of nuclear spin diffusion. The evolution of the polarization  $p_j(t) = \langle I_j^z(t) \rangle$  of the  $j$ th nucleus, placed at site  $\mathbf{r}_j$ , and having the spin  $I$ , is described by the kinetic equation

$$\dot{p}_j = - \sum_{a=1}^N v_{aj} p_j = - \sum_{\mathbf{r}} n_{\mathbf{r}} v_{\mathbf{r}j} p_j, \quad p_j(t=0) = 1. \quad (1)$$

Here  $v_{aj}$  is the depolarization rate under influence of the  $a$ th paramagnetic center (acceptor),  $N$  is the total number of acceptors in the sample,  $v_{\mathbf{r}j} = v_{aj}(\mathbf{r}_a = \mathbf{r})$ , and  $n_{\mathbf{r}}$  is the occupation number of the site  $\mathbf{r}$  by an acceptor ( $n_{\mathbf{r}} = 1$  (0) if the site  $\mathbf{r}$  is (not) occupied by an acceptor). Occupations of different sites will be assumed to be independent and having no dependence on  $\mathbf{r}$  (with a small probability of occupation  $c \ll 1$  as a rule):

$$\langle n_{\mathbf{r}} \rangle_c = c, \quad \langle n_{\mathbf{r}} n_{\mathbf{x}} \rangle_c = c \delta_{\mathbf{r}\mathbf{x}} + c^2(1 - \delta_{\mathbf{r}\mathbf{x}}), \\ \left\langle \prod_{j=1}^m n_{\mathbf{r}_j} \right\rangle_c = \prod_{j=1}^m \langle n_{\mathbf{r}_j} \rangle_c = c^m. \quad (2)$$

All  $\mathbf{r}_j$  are different in the last relation. Coincidence in indexes can be treated using the identity  $n_{\mathbf{r}}^2 = n_{\mathbf{r}}$ . The problem consists of calculation of the observable polarization, averaged over all possible positions of acceptors in the sample. Ensemble averaging can be used for macroscopic samples:  $p(t) = \langle p_j(t) \rangle_c$ . Occupation number representation gives the simplest and general solution of the last

problem [5]. Indeed

$$\begin{aligned} p(t) &= \langle p_j(t) \rangle_c = \left\langle \exp \left( - \sum_{\mathbf{r}} n_{\mathbf{r}} v_{\mathbf{r}j} t \right) \right\rangle_c \\ &= \prod_{\mathbf{r}} \langle \exp(-n_{\mathbf{r}} v_{\mathbf{r}j} t) \rangle_c \\ &= \prod_{\mathbf{r}} \langle 1 + n_{\mathbf{r}} [\exp(-v_{\mathbf{r}j} t) - 1] \rangle_c \\ &= \prod_{\mathbf{r}} \langle 1 + c [\exp(-v_{\mathbf{r}j} t) - 1] \rangle_c \\ &= \exp \left\{ \sum_{\mathbf{r}} \ln [1 + c (\exp(-v_{\mathbf{r}j} t) - 1)] \right\}. \quad (3) \end{aligned}$$

The identity  $f(n_{\mathbf{r}}) = f(0) + n_{\mathbf{r}}(f(1) - f(0))$  is applied here. It is valid for any realistic function  $f(x)$ . The relation (3) is exact for any  $c$ .

The typical depolarization rate is of the form  $v_{aj} = v_0 r_0^6 \chi(\mathbf{n}_{aj})/r_{aj}^6$ , where  $v_0$  scales the transport at minimal distance  $r_0$ ,  $\mathbf{n}_{aj} = \mathbf{r}_{aj}/r_{aj}$ ,  $\mathbf{r}_{aj} = \mathbf{r}_a - \mathbf{r}_j$ ,  $\chi(\mathbf{n}) = |Y_{21}(\mathbf{n})|^2$ , and  $Y_{lm}(\mathbf{n})$  is the spherical harmonics. Additional simplification is possible for the continuum media approximation (CMA), when  $c \rightarrow 0$ , but  $p(0) - p(t) \neq 0$ :

$$\begin{aligned} p(t, c \rightarrow 0) &= \exp \left( n \int d^d r (e^{-v_{\mathbf{r}j} t} - 1) \right) \\ &= \exp \left( - (\beta_F t)^{d/6} \right), \quad \beta_F \propto n^{6/d} v_0 r_0^6. \quad (4) \end{aligned}$$

Here arbitrary spatial dimension  $d$  is considered, and impurity density  $n = c/\Omega_c$  is introduced together with the prime cell volume  $\Omega_c$ .

The most important properties of Forster's result (4) consist of replacement of the simple exponential kinetics by a slower law, and in proportionality of  $\beta_F$  to transfer rate at average distance  $\bar{r} = n^{-1/d}$ :  $\beta_F = \text{const} \cdot v_{\mathbf{r}j}(|\mathbf{r} - \mathbf{r}_j| = \bar{r})$ , where const doesn't depend on concentration.

Anderson introduced the model to calculate the EPR line form function  $g(\omega) = \int_{-\infty}^{\infty} \frac{dt}{2\pi} e^{-i\omega t} F(t)$  and free induction decay  $F(t) = \langle \langle S_+(t) S_- \rangle \rangle_c / \langle \langle S_+ S_- \rangle \rangle_c$  for the system of equivalent spins  $S_j = 1/2$ , randomly distributed inside a sample and having dipole–dipole interaction. The model consists of substitution of exact secular interaction

$$\begin{aligned} H &= \frac{1}{2} \sum_{ij} b_{ij} \left( S_i^z S_j^z - \frac{1}{3} \mathbf{S}_i \mathbf{S}_j \right) \\ &= \frac{1}{2} \sum_{\mathbf{r}, \mathbf{q}} n_{\mathbf{q}} n_{\mathbf{r}} b_{\mathbf{r}\mathbf{q}} \left( S_{\mathbf{r}}^z S_{\mathbf{q}}^z - \frac{1}{3} \mathbf{S}_{\mathbf{r}} \mathbf{S}_{\mathbf{q}} \right) \end{aligned} \quad (5)$$

by a simpler one, depending on spin  $z$ -components only:

$$H_A = \frac{1}{2} \sum_{\mathbf{r}, \mathbf{q}} n_{\mathbf{q}} n_{\mathbf{r}} b_{\mathbf{r}\mathbf{q}} S_{\mathbf{r}}^z S_{\mathbf{q}}^z, \quad (6)$$

where  $S_{\mathbf{r}}^{\alpha}$  is the  $\alpha$  component of spin  $\mathbf{S}_{\mathbf{r}}$  placed at site  $\mathbf{r}$  and  $b_{\mathbf{r}\mathbf{q}} = b_0 r_0^3 (1 - 3 \cos^2 \vartheta_{\mathbf{r}\mathbf{q}}) / |\mathbf{r} - \mathbf{q}|^3$  is the standard

dipole–dipole coefficient [2] (here and below  $b_{rr} = 0$ ). It is of great importance, that free induction signals  $F_2(t) = \langle S_+(t)S_- \rangle_0 / \langle S_+S_- \rangle_0$  are the same for both Hamiltonians for two spin problem, if  $S_j = 1/2$ . Anderson's model has an exact solution for arbitrary concentration [6]. Indeed, the equation of motion for orthogonal spin components has the solution

$$\begin{aligned} n_r S_r^+(t) &= n_r \exp(iH_A t) S_r^+ \exp(-iH_A t) \\ &= n_r S_r^+ \exp\left(i \sum_{\mathbf{q}} n_{\mathbf{q}} b_{r\mathbf{q}} S_{\mathbf{q}}^z t\right). \end{aligned}$$

Introducing total moment of the sample  $\mathbf{S} = \sum_{\mathbf{r}} n_{\mathbf{r}} \mathbf{S}_{\mathbf{r}}$ , and using standard high-temperature approximation for free induction decay (FID) we have

$$\begin{aligned} F(t) &= \frac{\langle \langle S_+(t)S_- \rangle \rangle_c}{\langle \langle S_+S_- \rangle \rangle_c} = \left\langle \left\langle \exp\left(i \sum_{\mathbf{q}} n_{\mathbf{q}} b_{r\mathbf{q}} S_{\mathbf{q}}^z t\right) \right\rangle \right\rangle_c \\ &= \prod_{\mathbf{q}} \langle \cos(n_{\mathbf{q}} b_{\mathbf{q}r} t/2) \rangle_c \\ &= \prod_{\mathbf{q}} (1 + c (\cos(b_{\mathbf{q}r} t/2) - 1)) \\ &= \exp\left(\sum_{\mathbf{q}} \ln(1 + c (\cos(b_{\mathbf{q}r} t/2) - 1))\right). \quad (7) \end{aligned}$$

Applying CMA to (7) we arrive at Anderson's result:

$$\begin{aligned} F(t, c \rightarrow 0) &= \exp\left(-n \int d^d q (1 - \cos(b_{\mathbf{q}r} t/2))\right) \\ &= \exp\left(-(D_A t)^{d/3}\right). \quad (8) \end{aligned}$$

For three-dimensional systems the model evidently has simple Lorentz form function:

$$g(\omega) = g_A(\omega) = \frac{D_A}{\pi (\omega^2 + D_A^2)}.$$

Anderson's parameter  $D_A$ , as well as  $\beta_F$ , is proportional to the rate of the process taken at the average distance:  $D_A \propto b_{\mathbf{q}r}(|\mathbf{q} - \mathbf{r}| = \bar{r})$ . Other derivations of Eq. (8) (closer to the original Anderson's treatment fulfilled for  $d = 3$ ) can be found in [7].

### Delocalization of Nuclear Polarization in a Disordered Spin System

One of the most important generalizations of the Forster process is presented by delocalization of nuclear polarization in the system of impurity nuclei, randomly distributed in a diamagnetic matrix, when host nuclei have

faster phase relaxation and flip-flop transitions than impurities. Examples of such systems: nuclei  $^6\text{Li}$  in the single crystal  $^7\text{Li}^{19}\text{F}$  or spins  $^{107}\text{Ag}$  in the single crystal  $^{109}\text{Ag}^{19}\text{F}$ . The former system is of special interest, because it is accessible for direct experimental study due to unique coincidence of  $g$ -factors of stable nuclei  $^6\text{Li}$  and  $\beta$ -active nuclei  $^8\text{Li}$  ( $\beta$ -nuclei) [8]. The system has simple and instructive ergodic properties as well [9]. Therefore, we will consider below a specific system (consisting of  $^6\text{Li}$  nuclei in a LiF single crystal with addition of one  $\beta$ -nucleus  $^8\text{Li}$ ) where nuclear polarization transfers from initially polarized  $^8\text{Li}$  nucleus to the nearest nonpolarized  $^6\text{Li}$  nuclei and then migrates over other  $^6\text{Li}$  nuclei and might return back to the  $^8\text{Li}$ .

It was shown in [10] that modern master equation treatment leads to a description of the system by following kinetic equations, which have been formulated in [5]:

$$\frac{\partial p_{i0}}{\partial t} = - \sum_j (v_{ji} p_{i0} - v_{ij} p_{j0}), \quad p_{i0}(t=0) = \delta_{i0}. \quad (9)$$

Here  $p_{i0} = \langle I_i^z \rangle$  is the quantum statistical average value of the  $z$ -component (polarization) of the  $i$ th nucleus of the system  $^8\text{Li}$ - $^6\text{Li}$ , placed at  $\mathbf{r}_i$  ( $i = 0$  corresponds to  $^8\text{Li}$  with a spin  $I_0 = I = 2$ , and  $i \neq 0$  to  $^6\text{Li}$  with a spin  $I_i = S = 1$ ). The rates of polarization transfer are of the form:

$$\begin{aligned} v_{ji} &= \xi_j v_{ji}^0 r_0^6 \cdot \left( \frac{1 - 3 \cos^2 \theta_{ji}}{r_{ji}^3} \right)^2, \\ v_{ji}^0 &= \frac{\pi}{6} S(S+1) \left( \frac{g_i g_j \beta_n^2}{\hbar r_0^3} \right)^2 g_{ij}(\omega_{ij}). \end{aligned} \quad (10)$$

Here  $\xi_j = I_j(I_j + 1)/[S(S + 1)]$ ,  $g_0 = g_I = 0.8267$  and  $g_{i \neq 0} = g_S = 0.8220$  are  $g$ -factors of  $^8\text{Li}$  and  $^6\text{Li}$  correspondingly,  $\beta_n$  is nuclear magneton,  $\theta_{ji}$  is the angle between external static field  $H_0$  and  $\mathbf{r}_{ji} = \mathbf{r}_j - \mathbf{r}_i$ ,  $r_0 = 2.01 \cdot \sqrt{2} \text{ \AA}$  is the minimal distance between Li nuclei, and  $\omega_{ij}$  is the difference of the Larmor frequencies. The cross-relaxation form function  $g_{ij}(\omega)$  was studied in [11], and it can be taken as Gaussian  $g_{ij}(\omega) \approx \exp(-\omega^2/(2M))/(2\pi M)^{1/2}$  in the main approximation. Here  $M = 2M_2$ , where  $M_2$  is the second moment of the  $^8\text{Li}$  NMR line. As a result  $v_{ij}^0$  have two values only,  $v_{ij}^0 = v_0$  for transfer between  $^6\text{Li}$  spins (at that  $\omega_{ij} = 0$ ), and  $v_{i0}^0 = v_1$  for transfer between  $^8\text{Li}$  and  $^6\text{Li}$  (with  $\omega_{ij} = \Delta$ ).



The occupation number representation of the Eq. (9) is of the form [5]

$$\frac{\partial \tilde{P}_{\mathbf{x}0}}{\partial t} = - \sum_{\mathbf{z}} (n_{\mathbf{z}} v_{\mathbf{zx}} \tilde{P}_{\mathbf{x}0} - n_{\mathbf{x}} v_{\mathbf{xz}} \tilde{P}_{\mathbf{z}0}) ,$$

$$\tilde{P}_{\mathbf{x}0}(t=0) = n_{\mathbf{x}} \delta_{\mathbf{x}0}/c , \quad (11)$$

where propagator  $\tilde{P}_{\mathbf{x}0}$  gives polarization of the lattice site  $\mathbf{x}$  when initially the site  $\mathbf{0}$  was polarized, and  $v_{\mathbf{zx}} = v_{ij}(\mathbf{r}_i = \mathbf{z}, \mathbf{r}_j = \mathbf{x})$ . Equation (9) is a direct consequence of Eq. (11), that is evident, if we omit all empty sites for which  $n_{\mathbf{x}} = 0$  and, consequently,  $\tilde{P}_{\mathbf{x}0} = 0$ .

The experimentally observable value is polarization of the  $\beta$ -nucleus, averaged over a random distribution of  $^8\text{Li}$ - $^6\text{Li}$  nuclei  $P_{00}(t) = \langle p_{00} \rangle_c = \langle \tilde{P}_{00} \rangle_c$ . A calculation of such a value belongs to the problems of random walks in disordered media (RWDM), which is one of the most complex modern fields of statistical physics, which defines the actuality of this kind of study. To clarify the status of the problem we note that its solution in CMA can be connected [12] with the calculation of the path integral of the form:

$$P_{\mathbf{xy}}(t) = \int_{\mathbf{q}(0)=\mathbf{x}}^{\mathbf{q}(t)=\mathbf{y}} D\mathbf{p}(\tau) D\mathbf{q}(\tau) \exp \left[ i \int_{\mathbf{x}}^{\mathbf{y}} \mathbf{p} d\mathbf{q} + L[\mathbf{q}, \mathbf{p}, t] \right] \quad (12)$$

$$L[\mathbf{q}, \mathbf{p}, t] = n \int d^3z \left( e^{-\int_0^t d\tau A^z(\mathbf{q}(\tau), \mathbf{p}(\tau))} - 1 \right) , \quad (13)$$

where  $A^z(\mathbf{q}, \mathbf{p}) = v_{\mathbf{zq}}(1 - e^{-i\mathbf{p}(\mathbf{z}-\mathbf{q})})$ . The representation (12) is similar to, but more complex than path integrals in the famous polaron problems (PP) [12]. The main difference consists of multi-time action  $L$  in (13) instead of two-time action in PP, in strong singularity of  $v_{\mathbf{zq}}$  instead of the less singular kernel  $1/|\mathbf{z}-\mathbf{q}|$  in PP, in additional path integral over all  $\mathbf{p}(\tau)$  and in strong dependence of the infinite-fold integral (12) on the exact form of approximations by the integrals with finite multiplicity.

Field and superfield path integral representations for  $P_{\mathbf{xy}}(t) = \langle \tilde{P}_{\mathbf{xy}} \rangle_c$  exist as well [13]. They demonstrate the relation of the RWDM with general problems of the modern field theory. It should be noted, that Eq. (9) and corresponding propagator  $P_{\mathbf{xy}}(t)$  have wide applications in many fields of physics. For example they describe (after minimal corrections, but retaining the dipole long-range action) incoherent spatial transport of the localized electronic excitations (which is of importance in optics and biophysics) [14], and with  $\ln(v_{ij}) \propto r_{ij}$  the same equations are used in the theory of the hopping conductivity [15], where their applications are combined with percolation theory.

The prognosis of the measurements with dipole transport has been based on the relation [10]

$$P_{00}(t) = F(t) = \exp(-\sqrt{\beta_1 t}) + \xi \frac{1 - \exp(-\sqrt{\beta_1 t})}{(\mu\beta(t+\tau))^{3/2}} \cdot \left( 1 + \frac{\varphi}{\sqrt{\mu\beta(t+\tau)}} \right) , \quad (14)$$

where  $\xi = \xi_0 = I(I+1)/[S(S+1)] = 3$ , Forster parameters  $\beta = \frac{256}{243} \frac{r_0^6}{\Omega_c^2} \pi^3 c^2 v_0$  and  $\beta_1 = \beta \cdot v_1/v_0$ , and the limit of small  $c$  is assumed (CMA). The values  $\varphi = 2.09$  and  $\mu\beta\tau = 5.11$  were chosen here to construct Eq. (14) as an interpolating formula between exact results [16] of the expansion of  $P_{00}$  in terms of  $c^m$  (which is an expansion in powers of  $(\beta t)^{m/2}$  in reality) and expansion in terms of  $1/(\beta t)^{m/2}$ , produced by the approximative or numerical treatment of the long time asymptotics (where dipole long ranging induces exact dependence between the first and second terms [10]). The value of  $\mu$  is connected with the main values  $D_\alpha$  of the diffusion tensor:  $\mu\beta = 4\pi(c/\Omega_c)^{2/3} (\prod_{\alpha=1}^3 D_\alpha)^{1/3}$ . It should be stressed that an analytical solution for the nature of the long time asymptotics is absent up to now, but studies of realistic models [17] and modern numerical-analytical studies [13,18,19] indicate that it is of diffusion type. The diffusion tensor is calculated now with 1% uncertainty [18,19] giving  $\mu = 0.71$ . As a result the relation (14) holds to within  $(\beta t)^{1/2}$  at small  $\beta t$ , and it holds to within  $(\beta t)^{-2}$  at large  $\beta t$ .

The applicability of the Eq. (14) (having no fitting parameters) was checked at  $\beta_1 t \leq 10$  in [8] and at  $\beta_1 t \leq 15$  in [20]. The last experimental results [21,22] of the ITEP group indicate that at  $\beta_1 t \sim 25$  some correction is necessary. It can be introduced in simplest form as

$$P_{00}(t) = F(t)G(t)$$

$$G(t) = G_{\text{exp}} = \left( 1 - \frac{(\frac{1}{8} + \alpha) \beta_1 t - u(\beta_1 t)^2}{(1 + \nu\beta t)^3} \right) . \quad (15)$$

Here  $F(t)$  is defined in (14),  $\alpha = \alpha(\Delta)$  is tabulated in [10] (at that  $\alpha(\Delta \rightarrow 0) = 0.013$ ), and fitting parameters  $u$  and  $\nu$  should be determined by experimental data. The relation (15) holds to within  $\beta t$  at small  $\beta t$ , and it holds to within  $(\beta t)^{-2}$  at large  $\beta t$ . The fitting produces  $u \approx 0.06$  and  $\nu \approx 0.12$  [21]. Direct numerical simulation of the  $P_{00}(t)$  fulfilled in [19] for a small external magnetic field corresponding to  $\beta \approx \beta_1$  justified this correction and gave more precise form and values for  $F(t)$  indicating that

$0.9 \leq F(t) \leq 2$ . New experimental results [23] have been obtained for larger magnetic fields, when  $\beta - \beta_1 \propto \beta_1$ . More detailed numerical simulation is expected for description of the process in this region of fields. This should take into account correlations of local fields on impurity spins [11].

The numerical simulation [13,18,19] is based on the substitution of infinite disordered media by a crystal with a large disordered elementary cell, containing  $N_d \sim 1000$  impurity spins, and on checking of stability of the results with regard to variation of  $N_d$ . It starts from Eq. (9) written as

$$\frac{\partial p_{i0}}{\partial t} = - \sum_j B_{ij} p_{j0}, \quad B_{ij} = \delta_{ij} \sum_k v_{ki} - v_{ij},$$

$$p_{i0}(t=0) = \delta_{i0}, \quad (16)$$

and  $N_d$  impurities randomly placed in sites of a supercell, having  $N = N_d/c = N_g^3$  lattice sites, and edges  $\mathbf{R}_\alpha = N_g \mathbf{b}_\alpha$ , where  $\mathbf{b}_\alpha$  form the basis of the matrix crystal (LiF for the system  $^8\text{Li-}^6\text{Li}$ ). Then the supercell is continued periodically to cover all space. If  $N_d \rightarrow \infty$ , then we go back to random media. It is very important that the initial condition remains of the correct form and that it is not continued periodically. The eigenvalue problem

$$\sum_j B_{ij} \phi_j(m) = \varepsilon_m \phi_i(m) \quad (17)$$

has Bloch's solution

$$\phi_j(m) = \exp(i\mathbf{k}\mathbf{r}_j) \chi_j(\mathbf{k}, \mu), \quad m = \{\mathbf{k}, \mu\}, \quad (18)$$

where  $\chi_j(\mathbf{k}, \mu)$  has periodical dependence on impurity position  $\mathbf{r}_j$  and  $\mathbf{k}$  belongs to Brillouin zone  $V_B$  formed by all  $\mathbf{k}$  satisfying the condition

$$|\mathbf{k}\mathbf{R}_\alpha| \leq \pi.$$

Therefore, we have a finite eigenproblem for  $\chi_j(\mathbf{k}, \mu)$ :

$$\sum_{j=0}^{N_d-1} \bar{B}_{ij}(\mathbf{k}) \chi_j(\mathbf{k}, \mu) = \varepsilon_{\mathbf{k}, \mu} \chi_j(\mathbf{k}, \mu), \quad (19)$$

$$\bar{B}_{ij}(\mathbf{k}) = \sum_{n_\alpha} e^{-i\mathbf{k}(\mathbf{r}_i - \mathbf{r}_j - \mathbf{R}(n))} B_{ij}(\mathbf{r}_j + \mathbf{R}(n)). \quad (20)$$

Here  $\mathbf{R}(n) = \sum_{\alpha=1}^3 n_\alpha \mathbf{R}_\alpha$ , and  $B_{ij}(\mathbf{r}_j + \mathbf{R}(n))$  is  $B_{ij}$  with substitution  $\mathbf{r}_j \rightarrow \mathbf{r}_j + \mathbf{R}(n)$ . As a result

$$p_{j0}(t) = \frac{1}{V_B} \int_{V_B} d^3k \exp(i\mathbf{k}\mathbf{r}_j) (\exp(-\bar{B}(\mathbf{k})t))_{j0}, \quad (21)$$

$$P_{00}(t) = \frac{1}{V_B} \int_{V_B} d^3k \langle (\exp(-\bar{B}(\mathbf{k})t))_{00} \rangle_c, \quad (22)$$

$$P_{x0}(t) = \frac{c}{V_B} \int_{V_B} d^3k \exp(i\mathbf{k}\mathbf{x}) \langle (\exp(-\bar{B}(\mathbf{k})t))_{10} \rangle_c^{(r_1=\mathbf{x})}. \quad (23)$$

Here  $\langle (\exp(-\bar{B}(\mathbf{k})t))_{10} \rangle_c^{(r_1=\mathbf{x})}$  means that averaging is fulfilled under condition that site  $\mathbf{x}$  is occupied by an impurity spin having number 1 (the site  $\mathbf{0}$  is always occupied by spin "0").

For identical impurities

$$P_{00}(t) = \frac{1}{V_B N_d} \int_{V_B} d^3k \langle \text{Tr} \{ \exp(-\bar{B}(\mathbf{k})t) \} \rangle_c. \quad (24)$$

The Eqs. (22–24) were applied for numerical simulation basing on modern programs of matrix diagonalization. The results for the  $^8\text{Li-}^6\text{Li}$  system and for electrodipole transport of localized electronic excitations among identical impurities can be found in [19].

### Nuclear Relaxation via Paramagnetic Impurities

Another important generalization of Forster's process is presented by nuclear relaxation via paramagnetic impurities, which is the main relaxation channel in isolators if nuclear spin  $I = 1/2$ . The fundamental studies [2,24,25,26] were based on calculation of the linear (in impurity concentration  $c$ ) term of the time dependence of the sample magnetization  $P(t) = 1 - cQ(t)$  (for 3d-systems) with following substitution  $P(t) = \exp(-cQ(t))$ . Analysis of two- and one-dimensional problems was absent whereas experiments are already aimed at fractal objects [27]. Therefore, new theory was constructed [28]. It produced the relation  $P(t) = \exp(-cQ(t))$  as the main approximation, and the function  $Q(t)$  is calculated for arbitrary  $d \leq 3$ .

The process used to be described by the kinetic equation

$$\frac{\partial p(\mathbf{x}, t)}{\partial t} = D \Delta p(\mathbf{x}, t) - \sum_{\mathbf{z}} n_{\mathbf{z}} v_{\mathbf{zx}} p(\mathbf{x}, t),$$

$$v_{\mathbf{zx}} = \frac{v_0 r_0^6}{|\mathbf{x} - \mathbf{z}|^6} = \frac{C}{|\mathbf{x} - \mathbf{z}|^6}, \quad (25)$$

with the initial condition  $p(\mathbf{x}, t) = p_0$ . Here  $p(\mathbf{x}, t)$  is polarization of the nucleus, placed at the crystal site  $\mathbf{x}$ ,  $D$  is spin-diffusion coefficient, and the angular dependence of  $v_{\mathbf{zx}}$  is neglected together with the difference in eigenvalues of the diffusion tensor. The observable nuclear polariza-

tion (normalized to  $\bar{p}(t = 0) = 1$ ) is

$$\bar{p}(t) = \frac{1}{\Omega} \int d^d x p(\mathbf{x}, t) = \langle 0 | G(t) | 0 \rangle = \langle 0 | \langle G(t) \rangle_c | 0 \rangle. \quad (26)$$

Here  $d$  is the space dimensionality,  $\Omega$  is the crystal volume, the symbol  $|0\rangle$  presents a vector having components  $\langle \mathbf{x} | 0 \rangle = 1/\sqrt{\Omega}$ , and the propagator  $G_{\mathbf{xy}}(t) = \langle \mathbf{x} | G(t) | \mathbf{y} \rangle$  obey Eq. (25), but for initial condition  $G_{\mathbf{xy}}(t = 0) = \delta(\mathbf{x} - \mathbf{y})$ .

Expansion of the observable  $\bar{p}(t)$  in concentration powers [16] gives in first terms

$$\begin{aligned} \bar{p}(t) &= \langle 0 | G(t) | 0 \rangle = \langle 0 | G^{(0)}(t) | 0 \rangle \\ &+ n \int d^d r \langle 0 | [G^{(1)}(t, \mathbf{r}) - G^{(0)}(t)] | 0 \rangle + O(n^2) \\ &= \exp(-M_0(t)(1 + O(n^2))), \end{aligned} \quad (27)$$

$$M_0(t) = n \int d^d r \langle 0 | [G^{(0)}(t) - G^{(1)}(t, \mathbf{r})] | 0 \rangle. \quad (28)$$

The propagator  $G^{(0)}(t)$  corresponds to evolution in the absence of acceptors, and  $G^{(1)}(t, \mathbf{r})$  is the propagator of the system, having only one acceptor, placed at  $\mathbf{r}$ . Then, using operator representation of Eq. (25), translational invariance, resolvent identity and spectral expansion, we have

$$\begin{aligned} M_0(\lambda) &= \int_0^\infty dt e^{-\lambda t} M_0(t) \\ &= \frac{n\Omega}{\lambda} \left\langle 0 \left| U_0 \frac{1}{\lambda + A + U_0} \right| 0 \right\rangle \\ &= \frac{n\Omega}{\lambda} \left\langle 0 \left| U_0 \frac{1}{(A + U_0)(\lambda + A + U_0)} U_0 \right| 0 \right\rangle \\ &= \frac{n\Omega}{\lambda} \sum_n \frac{|\langle n | U_0 | 0 \rangle|^2}{(\lambda + E_n) E_n}. \end{aligned} \quad (29)$$

Here  $A = -D\Delta$ ,  $\langle \mathbf{x} | U_0 | \mathbf{z} \rangle = \delta_{\mathbf{xz}} v_{\mathbf{x0}}$ , and  $(A + U_0) | n \rangle = E_n | n \rangle$ . This representation is too complex for direct calculations, but it is useful for different asymptotical treatments (see, for example, [8]). The most important are short time and long time studies, because a satisfactory precision can be achieved with the aid of the representation

$$M_0(t) = M_F(t) + M_1(t). \quad (30)$$

Here the first term

$$M_F(t) = n \int d^d x (1 - \exp(-v_{\mathbf{x0}} t)) = (\beta_F t)^{d/s}$$

describes the initial (Forster's) part of the relaxation, while  $M_1(t)$  is the long-time asymptotic expression, calculated on the basis of the representation (30) and formally continued to arbitrary positive values of  $t$ . Here we use  $v_{\mathbf{x0}} = v_0 r_0^s / x^s$  to clarify some parametrical dependencies. The results, obtained for  $M_1(t)$  at arbitrary  $d \leq 3$ , are presented in detail in Ref [28]. In particular, at long time

$$\begin{aligned} M_1(t, d = 1) &= 4n \sqrt{\frac{Dt}{\pi}}, \\ M_1(t, d = 2) &= \frac{4\pi D n t}{\ln(Dt/b^2)}, \\ M_1(t, d = 3) &= 4\pi D b n t, \end{aligned} \quad (31)$$

where  $b \propto (C/D)^{1/(s-2)}$  is a “scattering length”, which incorporates all dependence on the “potential”  $v_{\mathbf{x0}}$ . It is evident, that at  $d = 1$  the dependence on  $b$  is absent here, and for  $d = 2$  it is rather weak.

In order to find an argument for regrouping (27) of the concentration expansion, we represent the propagator  $G(t)$  in the form

$$G(t) = \langle \exp(-(A + U)t) \rangle_c = \exp(-B(t)), \quad (32)$$

$$B(t) = At + M(t),$$

$$U = \sum_{\mathbf{z}} n_{\mathbf{z}} U^{\mathbf{z}}, \quad U_{\mathbf{xq}}^{\mathbf{z}} = \delta_{\mathbf{xq}} v_{\mathbf{xz}}. \quad (33)$$

The operator  $M(t)$ , which is as-yet undefined, can be written in the form  $M(t) = \sum_{\mathbf{z}} M_{\mathbf{z}}(t)$  similar to the form of  $U$  in (33). It can be said that the operators  $M_{\mathbf{z}}(t)$  must adequately describe the effect of acceptors in the so-called effective medium that appears upon averaging over the configurations of acceptors. It is therefore natural to assume that, on average, the propagator  $G(t)$  undergoes no changes if one of the sites of the effective medium is replaced by an actual one and if the result is thereupon averaged over the distribution of acceptors; that is,

$$G(t) = \langle \exp(-At - M(t) + M^z(t) - n_z U^z t) \rangle_c. \quad (34)$$

Relations (32)–(34) form a closed set of nonlinear operator equations. The solution, according to [28], practically coincides with  $P(t) = \exp(-M_0(t))$ , if  $cQ(t) = M_0(t) \sim 1$  and  $c \ll 1$ , and corrections are important for longer time.

It should be stressed, that (1) to clarify the influence of corrections to this solution in more detail we can calculate the next ( $\propto c^2$ ) term of the concentration expansion, and (2) there exist physical [29] and mathematical [30,31] studies, giving the law  $\log(1/\bar{p}(t \rightarrow \infty)) \propto t^{d/(d+2)}$ , which is expected [32] to be valid at  $\bar{p}(t) \lesssim 10^{-12}$ . The asymptotics can be received by methods of field theory, but up to a preexponential multiplier only [33], therefore these

methods did not produce useful results for the problem of random walks in disordered media, discussed in Sect. “[Delocalization of Nuclear Polarization in a Disordered Spin System](#)”, where the exponential part of the asymptotics is absent.

### Resonance Line Form Function for Magnetically Diluted Solids

The line shape and the Fourier-transform-related free induction decay (FID) belong to the most important observable values in the physics of magnetic resonance. In the study of nuclear spin systems forming a crystal lattice the first ( $\propto t^2$  and  $\propto t^4$ ) terms of the expansion of FID in powers of time carry information of high importance [2]. In the theory of the line shape of disordered (magnetically diluted) electron spin systems, the first ( $\propto c$  and  $\propto c^2$ ) terms in the expansion in powers of the concentration  $c$  of paramagnetic centers play the same role [6]. The third ( $\propto c^2$ ) term was calculated in a recent study [34] for the first time. This consideration is particularly topical in connection with the new experiments on measuring the EPR spectra of paramagnetic impurities distributed at the solid surface [35,36]. It can be expected, that modern pulse methods will produce new possibilities for measurement of FID in such magnetically disordered nuclear systems as  $^{29}\text{Si}$  in silicon crystals for example. Some efforts in this direction can be found in the recent article [37] and in references therein.

Let the paramagnetic centers (PCs) be randomly distributed in a  $d$ -dimensional crystal lattice with the prime-cell volume  $\Omega_c$ . The free induction decay in the high-temperature approximation is given by

$$G(t) = \langle \langle S^+(t) S^- \rangle_0 \rangle_c / \langle \langle S^+ S^- \rangle_0 \rangle_c, \quad (35)$$

where  $S^\pm = \sum_{\mathbf{r}} n_{\mathbf{r}} S_{\mathbf{r}}^\pm$ ,  $S_{\mathbf{r}}^\pm = S_{\mathbf{r}}^x \pm i S_{\mathbf{r}}^y$ ,  $S^+(t) = e^{iHt} S^+ e^{-iHt}$ ,  $n_{\mathbf{r}}$  is the occupation number,  $\langle \cdots \rangle_0 = \text{Tr}(\cdots) / \text{Tr}1$ ,  $\langle \cdots \rangle_c$  stands for the averaging over the spatial spin distributions (over occupation numbers), and  $H$  is the secular part of dipole-dipole interactions:

$$H = \frac{3}{4} \sum_{\mathbf{r}\mathbf{q}} n_{\mathbf{r}} n_{\mathbf{q}} A(\mathbf{r}, \mathbf{q}) \left( S_{\mathbf{r}}^z S_{\mathbf{q}}^z - \frac{a}{3} S_{\mathbf{r}} S_{\mathbf{q}} \right). \quad (36)$$

Here  $A(\mathbf{r}, \mathbf{q}) = \hbar \gamma^2 (1 - 3 \cos^2 \vartheta_{r\mathbf{q}}) / |\mathbf{r} - \mathbf{q}|^3$ ,  $\gamma$  is the gyromagnetic ratio, and  $\vartheta_{r\mathbf{q}}$  is the angle between  $\mathbf{r} - \mathbf{q}$  and external static field  $\mathbf{H}_0$ . Parameter  $a = 0$  in the Anderson model and  $a = 1$  for pure dipole interaction. For other  $a$  values, Hamiltonian (36) corresponds to a system with the anisotropic axisymmetric  $g$ -factor. In what follows,  $S = 1/2$  for all PCs. Let us expand FID in power of  $n_{\mathbf{x}}$  and

perform configurational averaging [6]. To terms  $\propto O(c^3)$  one has

$$G(t) = 1 + c \sum_{\mathbf{r}_1} (2K_{01}(t) - 1) + \frac{c^2}{2} \sum_{\mathbf{r}_1 \neq \mathbf{r}_2} (2K_{012}(t) - 2K_{01}(t) - 2K_{02}(t) + 1), \quad (37)$$

and, applying CMA we have up to terms  $\propto O(n^3)$

$$G(t) = 1 + n \int d^d r_1 (2K_{01}(t) - 1) + \frac{n^2}{2} \int d^d r_1 d^d r_2 (2K_{012}(t) - 2K_{01}(t) - 2K_{02}(t) + 1), \quad (38)$$

where  $n = c/\Omega_c$  is the  $d$ -dimensional PC density, and

$$K_{01}(t) = \langle e^{iH_{01}t} S_0^+ e^{-iH_{01}t} (S_0^- + S_1^-) \rangle_0, \\ K_{012}(t) = \langle e^{iH_{012}t} S_0^+ e^{-iH_{012}t} (S_0^- + S_1^- + S_2^-) \rangle_0.$$

Here, according to Eq. (36),

$$H_{ij} = \frac{1}{2} A_{ij} (3S_i^z S_j^z - a \mathbf{S}_i \mathbf{S}_j), \quad H_{012} = H_{01} + H_{02} + H_{12}, \quad (39)$$

with  $A_{ij} = A(\mathbf{r}_i, \mathbf{r}_j)$ . The interaction  $A_{ij} \propto |\mathbf{r}_i - \mathbf{r}_j|^{-3}$ . Therefore, substitution of integration variables  $\mathbf{r}_i \rightarrow t^{1/3} \mathbf{r}_i$  excludes time from the integrands and reveals that the  $m$ th term in (38) is  $\propto (nt^{d/3})^m$ . That is Eq. (38) is expanded [16] in terms of dimensionless parameter

$$(D_d t)^{d/3} = n \int d^d r_1 (1 - 2K_{01}(t)), \quad (40)$$

and, hence, can be represented as

$$G(t) = 1 - (D_d t)^{d/3} + \frac{1}{2} \xi_d(a) (D_d t)^{2d/3} + O((D_d t)^d). \quad (41)$$

The functions  $\xi_d(a)$  were calculated numerically basing on the Eq. (38). It was supposed, that at  $d \leq 2$  all field directions are equally probable. For this case, after averaging of (41) we have

$$\bar{G}(t) = 1 - (\bar{D}_d t)^{d/3} + \frac{\bar{\xi}_d(a)}{2} (\bar{D}_d t)^{2d/3} + O((\bar{D}_d t)^d), \quad (42)$$

and

$$\bar{D}_d = \beta_d n^{d/3} \gamma^2 \hbar, \quad \beta_3 = \frac{2\pi^2}{3\sqrt{3}}, \quad (43) \\ \beta_2 = 4.647, \quad \beta_1 = 6.348.$$

It should be noticed, that  $\tilde{D}_3 = D_3$ , and  $D_3$  coincides with Anderson's result [4].

The results are presented in the following table:

|         |       |      |      |      |      |      |      |      |      |
|---------|-------|------|------|------|------|------|------|------|------|
| $a$     | 0     | 0.2  | 0.4  | 0.6  | 0.8  | 1.0  | 1.2  | 1.4  | 1.6  |
| $\xi_3$ | 1     | 1.01 | 1.03 | 1.05 | 1.08 | 1.11 | 1.13 | 1.15 | 1.18 |
| $\xi_2$ | 1.027 | 1.07 | 1.11 | 1.15 | 1.19 | 1.22 | 1.25 | 1.27 | 1.29 |
| $\xi_1$ | 1.062 | 1.16 | 1.20 | 1.25 | 1.28 | 1.32 | 1.35 | 1.37 | 1.40 |

(44)

These relations can be used directly for analysis of the wings of experimental EPR lines [34]. But treatment of the full line or FID requires regrouping of the expansion (42) in such a way to receive a physically adopted result for all  $t$ . One of the simplest approaches [38,39,40] was generalized for this aim in [34]. It introduces the most essential properties of the disordered systems into the Anderson–Weiss–Kubo (AWK) theory [41,42], which was originally developed for the description of line narrowing by motion. The result is of the form

$$G(t) = \exp \left( - \left( 2B_d^2 \int_0^t d\tau (t - \tau) F(B_d \tau) \right)^{d/6} \right), \quad (45)$$

$$F(x) = \exp \left( -(q_d x)^{d/3} \right).$$

Being approximative Eq. (45) nevertheless reproduces the structure of the expansion (41), non-negativity of the resonance line shape and parameters  $B_d$  and  $q_d$  can be defined using  $D_d$  and  $\xi_d$ , that gives

$$B_3 = D_3, \quad B_2 = \tilde{D}_2, \quad (46)$$

$$q_3 = 3(\xi_3 - 1), \quad q_2 = ((10/3)(\xi_2 - 1))^{3/2}.$$

The dimensionality  $d = 1$  requires more refined treatment, because system orientations near the “magic” direction  $\theta = \arccos(1/\sqrt{3})$  produce a singularity  $g_1(\omega \rightarrow 0) \propto \ln(1/\omega)$ .

In application to pure dipole interaction with  $a = 1$  the Eqs. (46) and (44) give  $q_3 = 0.33 \approx 1/3$  and  $q_2 = 0.63$ . Fitting of the experimental data of [35] in the region, where the sample was considered as two-dimensional gives  $q_2 = q_2^{\text{exp}} \approx 0.05$  with significant distinctions from both  $q_2 = 0$ , and  $q_2 = 0.63$ . Additional studies are necessary to clarify the nature of this deviation from pure dipole evolution.

Application of the approach to magnetically diluted nuclear spin systems (like  $^{29}\text{Si}$  in silicon) requires one important modification connected with the fact that typical nuclear concentration  $c = 0.01 \div 0.1$  is not sufficiently small to replace lattice sums by integrals in Eq. (38). There-

fore, FID for such concentrations can be represented by the Eq. (45) with substitution

$$(D_d t)^{d/3} \rightarrow \kappa(t) = c \sum_{\mathbf{r}_1} (1 - 2K_{01}(t)). \quad (47)$$

For example, if  $d = 3$  and  $a = 1$ , then Eqs. (45), (46) produce

$$G(t) = \exp \left( - (6(D_3 t + 3 \exp(-D_3 t/3) - 3))^{1/2} \right). \quad (48)$$

Applying the recipe (47) we receive the representation

$$G(t) = \exp \left( - (6(\kappa(t) + 3 \exp(-\kappa(t)/3) - 3))^{1/2} \right), \quad (49)$$

which is exact at small  $\kappa(t)$  up to  $O(\kappa^2(t))$  and coincides with (48) for  $\kappa(t) \gtrsim 1$ .

### Saturation on the Wing of a Dipole-Broadened EPR Line and Cluster Expansions

A bright demonstration of peculiarities of spin dynamics in disordered media is realized in the shape of the hole burned on the wing of the EPR line. The EPR spectrum in dipole-broadened  $3d$  solids has the Lorentz wing, but experimentation [43] revealed that the wing of the hole falls down exponentially contrary to expectations, that a dipole broadened line should be homogeneous. To solve this contradiction a new method of cluster expansions [39,40] was invented, which gives a constructive alternative to both concentration expansion, discussed in Sect. “Resonance Line Form Function for Magnetically Diluted Solids”, and spin pockets conception [44], used in [43] for interpretation of the data.

If interaction inside a pair of spins exceeds interactions with any other spin, then the spectrum of the pair is similar to a discrete one, broadened by small interactions with surrounding spins. The pair forms a two-spin cluster (2-cluster). Analogically we can define 3- and other many-spin clusters. It is sufficient for many tasks to divide all spins into three groups: 2- and 3-clusters and all other “mass” spins. Numerical analysis [39] shows, that 2- and 3-clusters contain 51(1)% and 11(1)% of all spins correspondingly, but 2-clusters define all Lorentz wings of the EPR line. Mass spins (similar to nuclear spins in a crystal) have finite heat capacity, which is defined by interaction at average distance contrary to infinite heat capacity of a full dipole–dipole reservoir, calculated in CMA. Mass spins produce fluctuating fields, which gave the main broadening of the clusters spectra. Transitions between states of the cluster take place due to interac-



tions with other clusters and with mass spins, and they are slow. As a result saturation at the EPR line wing induce transitions between states of the 2-cluster, and the line shape of the hole is defined by interaction of the cluster with mass spins. If this interaction is estimated within the AWK model, then any mass configuration produces exponential wings of the hole  $g(\Delta) \propto \exp(-\Delta/\mu)$ , where  $\Delta$  is detuning of the saturating field from the cluster transition frequency, and  $\mu$  is defined by magnitude of the fields produced by other spins on the cluster, and by the rate of their fluctuations. After configurational averaging the wings became  $\langle g(\Delta) \rangle_c \propto \Delta^{-4}$ , but averaged hole shape described the transitions for small magnitude  $\omega_1$  of the saturating field only. In a general case the observable area of the hole  $\sigma$  should be averaged. For large  $\omega_1$  we have  $\sigma \propto \ln(t_p \omega_1^2/\mu)$ , where  $t_p$  is the duration of the saturating pulse. Logarithm is a slow function, therefore  $\langle \sigma \rangle_c \propto \ln(t_p \omega_1^2/\mu)_c$ , which was observed in [43] and was interpreted as the exponential wing of Portis's pockets. The described theory produces a microscopic picture of the phenomenon and indicates that Portis's pockets have limited relation to the problem, because they are homogeneous by definition, i. e.  $g(\Delta) = \langle g(\Delta) \rangle_c$ , while in the cluster theory this relation isn't fulfilled.

### Future Developments

Methods described in Sect. "Delocalization of Nuclear Polarization in a Disordered Spin System" are expected to be useful for understanding of studies of delocalization of relaxed excitons in different problems which can be studied by optical methods (fluorescence and multi-wave mixing) first of all. We expect that results obtained in this section will also produce a reliable basis for development of general methods of statistical physics and path integration theory in order to reproduce and clarify these results and many other problems of nonlinear field theory.

The results of Sect. "Nuclear Relaxation via Paramagnetic Impurities" create a new regular method for treatment of a wide class of problems of disordered media. The method is similar to, but, nevertheless, different from standard coherent (effective) potential approximation. We expect that the same ideas applied to the second term of the concentration expansion can produce better understanding for many related problems.

The results of Sect. "Resonance Line Form Function for Magnetically Diluted Solids" require new experimental studies for all spatial dimensions. Checking on nuclear spin systems like  $^{29}\text{Si}$  in silicon crystals is most desirable, because the details of interaction of nuclei are known much better than for paramagnetic centers.

Cluster expansion, described in Sect. "Saturation on the Wing of a Dipole-Broadened EPR Line and Cluster Expansions" should be useful for different problems of spectral transport in disordered solids.

The work was supported by the Russian Foundation for Basic Research, project #06-02-17088.

### Bibliography

1. Balesku R (1975) Equilibrium and Non-Equilibrium Statistical Mechanics. Wiley, New York
2. Abragam A, Goldman M (1982) Nuclear Magnetism: Order and Disorder. Clarendon Press, Oxford
3. Forster T (1949) Z Naturforsch (A) 4:321
4. Anderson PW (1951) Phys Rev 82:342
5. Dzheparov FS, Lundin AA (1978) Sov Phys JETP 48:514
6. Dzheparov FS, Lundin AA, Khazanovich TN (1987) Sov Phys JETP 65:314
7. Feldman EB, Lacelle S (1996) J Chem Phys 104:2000
8. Abov YG, Bulgakov MI, Borovlev SP, Gul'ko AD, Garochkin VM, Dzheparov FS, Stepanov SV, Trostin SS, Shestopal VE (1991) Sov Phys JETP 72:534
9. Dzheparov FS (1999) JETP 89:753
10. Dzheparov FS (1991) Sov Phys JETP 72:546
11. Abov YG, Gulko AD, Dzheparov FS, Stepanov SV, Trostin SS (1995) Phys Part Nucl 26:692
12. Feynman RP (1972) Statistical mechanics. W.A. Benjamin, Massachusetts
13. Dzheparov FS, L'vov DV, Shestopal VE (1998) JETP 87:1179
14. Rieger PT, Palese SP, Miller RJD (1997) Chem Phys 221:85
15. Ziman JM (1979) Models of Disorder: the Theoretical Physics of Homogeneously Disordered Systems. Cambridge University Press, Cambridge
16. Dzheparov FS, Smelov VS, Shestopal VE (1980) JETP Lett 32:47
17. Dzheparov FS, Shestopal VE (1993) Theor Mat Phys 94:496
18. Dzheparov FS, L'vov DV, Shestopal VE (2007) J Supercond Nov Magn 20(2):175
19. Dzheparov FS (2005) JETP Lett 82(8):521
20. Dzheparov FS, Gul'ko A, Heitjans P, L'vov D, Schirmer A, Shestopal V, Stepanov S, Trostin S (2001) Physica B297:288
21. Abov YG, Gul'ko AD, Dzheparov FS (2006) Phys At Nuclei 69:1701
22. Gul'ko AD, Ermakov ON, Stepanov SV, Trostin SS (2007) J Supercond Nov Magn 20(2):169
23. Dzheparov FS, Gul'ko AD, Ermakov ON et al (2008) In: Salikhov KM (ed) Modern development of magnetic resonance, Abstracts of the International conference Zavoiyskiy-100. Kazan, pp 47–48, Appl Magn Reson 35(3)
24. Khutsishvili G (1965) Sov Phys Usp 8:743
25. Alexandrov IV (1975) Theory of Magnetic Relaxation. Nauka, Moscow (in Russian)
26. Atsarkin VA (1980) Dynamical Nuclear Polarization. Nauka, Moscow (in Russian)
27. Tabti T, Goldman M, Jacquinot JF, Fermon C, LeGoff G (1997) J Chem Phys 107:9239
28. Dzheparov FS, Jacquinot JF, Stepanov SV (2002) Phys At Nuclei 65:2052
29. Balagurov VY, Vaks VT (1973) Sov Phys JETP 38:968
30. Pastur LA (1977) Teor Mat Fiz 32:88

31. Donsker M, Varadhan S (1975) Commun Pure Appl Math 28:525
32. Havlin S, Dishon M, Kiefer JE, Weiss GH (1984) Phys Rev Lett 53:407
33. Renn SR (1986) Nucl Phys B275(FS17):273
34. Dzheparov FS, Kaganov IV (2002) JETP Lett 75:259
35. Atsarkin VA, Vasneva GA, Demidov VV, Dzheparov FS, Odintsov BM, Clarkson RB (2000) JETP Lett 72:369
36. Atsarkin VA, Demidov VV, Vasneva GA, Dzheparov FS, Ceroke PJ, Odintsov BM, Clarkson RB (2001) J Magn Res 149:85
37. Li D, Dementyev AE, Dong Y, Ramos R, Barrett SE (2007) Phys Rev Lett 98:190401
38. Grinberg ES, Kochelaev BI, Khaliullin GG (1981) Sov Phys Solid State 23:224
39. Dzheparov FS, Henner EK (1993) JETP 77:753
40. Dzheparov FS, Kaganov IV, Khenner EK (1997) JETP 85:325
41. Anderson PW, Weiss PR (1953) Rev Mod Phys 25:269
42. Kubo R (1962) J Phys Soc Jpn 17:1100
43. Atsarkin VA, Vasneva GA, Demidov VV (1986) Sov Phys JETP 64:898
44. Portis AM (1953) Phys Rev 91:1071

## Spin-Polarized Quantum Transport in Mesoscopic Conductors: Computational Concepts and Physical Phenomena

MICHAEL WIMMER, MATTHIAS SCHEID,  
KLAUS RICHTER  
Institut für Theoretische Physik, Universität Regensburg,  
Regensburg, Germany

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Numerical Quantum Transport  
Matrix Reordering Strategies for Quantum Transport  
Spin Filtering in Nanostructures  
Pure Spin Current Generation  
Future Directions  
Acknowledgments  
Bibliography

### Glossary

**Aharonov Bohm effect** The magnetic flux enclosed in between propagating quantum mechanical waves shifts their relative phases as a result of the underlying electromagnetic vector potential. This gives rise to distinct oscillations in the magnetoconductance of a ring conductor.

### Landauer–Büttiker formalism

For phase-coherent quantum transport, the Landauer–Büttiker formalism relates the conductance of a device to the transmission probability of charge carriers.

**Rashba- and Dresselhaus spin-orbit coupling** Coupling of the spin degree of freedom to the orbital motion of charge carriers due to structural or bulk inversion asymmetry in semiconductors.

**Ratchets** Devices that convert unbiased fluctuations or perturbations into directed motion.

**Spintronics** Extension of charge-based electronics in metals or semiconductors by utilizing the spin degree of freedom of the charge carriers.

### Definition of the Subject

Mesoscopic conductors are electronic systems of sizes in between nano- and micrometers, and often of reduced dimensionality. In the phase-coherent regime at low temperatures, the conductance of these devices is governed by quantum interference effects, such as the Aharonov–Bohm effect and conductance fluctuations as prominent examples. While first measurements of quantum charge transport date back to the 1980s, spin phenomena in mesoscopic transport have moved only recently into the focus of attention, as one branch of the field of spintronics. The interplay between quantum coherence with confinement-, disorder- or interaction-effects gives rise to a variety of unexpected spin phenomena in mesoscopic conductors and allows moreover to control and engineer the spin of the charge carriers: spin interference is often the basis for spin-valves, -filters, -switches or -pumps. Their underlying mechanisms may gain relevance on the way to possible future semiconductor-based spin devices.

A quantitative theoretical understanding of spin-dependent mesoscopic transport calls for developing efficient and flexible numerical algorithms, including matrix-reordering techniques within Green function approaches, which we will explain, review and employ.

### Introduction

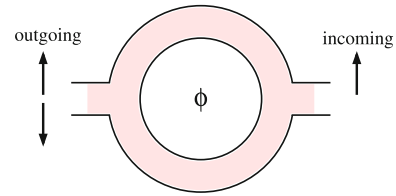
Charge and spin transport through phase-coherent conductors of mesoscopic scales carry imprints of wave interference as predominant and characteristic features: in the simplest case of a point contact, the conductance increases stepwise with Fermi energy, reflecting the discrete number of quantized open transverse channels contributing to transport; for more complex mesoscopic systems, such as ballistic quantum dots or diffusive conductors, the conductance typically wildly fluctuates, upon varying the Fermi energy or other parameters, around its classical

mean value. Among the different effects on charge transport, the Aharonov–Bohm (AB) effect represents possibly the most genuine interference phenomenon at the heart of mesoscopic physics: The magnetoconductance of a ring conductor coupled to two leads exhibits distinct sinusoidal oscillations when monitored as a function of a perpendicular magnetic field threading the ring, with a period given by the magnetic flux quantum. As the AB signal stems from interfering waves traveling through the two different arms of the ring, it requires phase coherent wave functions extending over the ring typically on micron scales [1]. Hence the AB effect is frequently being used as a tool to investigate phase coherence and dephasing mechanisms of the orbital part of the wave functions, while the spin degree of freedom was usually neglected.

With rising interest in spin-dependent transport, the interplay of the electron spin and charge degree of freedom has been exploited in a variety of spin interference devices, to be discussed below. Different types of couplings to the electron spin have been considered for spin engineering in non-magnetic conductors. On the one hand this is possible through the Zeeman coupling to an externally applied magnetic field. Non-uniform  $B$ -fields with spatially varying direction are being employed to achieve a tailored spin dynamics, including the possibility for guided spin evolution or triggering spin flips. On the other hand, intrinsic spin-orbit (SO) interaction proves to be relevant in spin-dependent transport. It exists in systems with bulk inversion asymmetry and/or structure inversion asymmetry, e. g. due to the vertical confinement in semiconductor heterostructures (Rashba SO coupling [2,3]).

Among the mesoscopic spin interference systems considered in the literature, ring geometries have again played an important role, opening up the field of spin-based AB physics (see [4,5] for recent accounts including overviews over the literature). This includes topics such as Berry phase-signatures in transport, both theoretically [6,7,8,9,10,11,12] and experimentally [13,14,15,16], spin-related conductance modulation [17,18,19], persistent currents [6,20], spin Hall effect [21], spin filters [22,23] and detectors [24], and spin switching mechanisms [4,10].

To illustrate how the spin polarization can be tuned by exploiting (orbital and spin) interference in such an AB setup, we consider as an introductory example spin switching in a two-dimensional (2D) ballistic phase-coherent ring symmetrically coupled to two single-channel leads (Fig. 1, [17]). We assume Rashba SO coupling which is relevant in conductors laterally defined on GaAs- or InAs-based two-dimensional electron gases (2DEGs). Rashba SO coupling will be defined and discussed in Sect. “Spin Filtering in Nanostructures”. It can be viewed

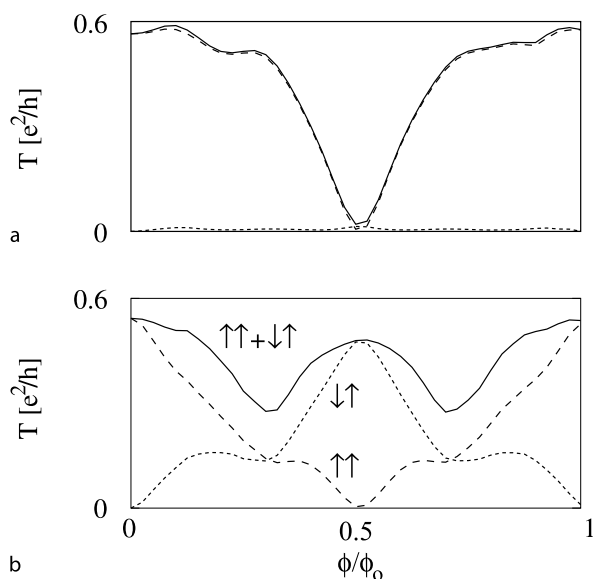


**Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 1**

**Aharonov–Bohm physics with spin:** Two-dimensional Aharonov–Bohm ring of mean radius  $r_0$  used for numerical calculations of the conductance presented in Fig. 2. An additional perpendicular magnetic field  $B$  generates a flux  $\phi = \pi r_0^2 B$ . The gray zone corresponds to the region subject to a finite Rashba coupling switched adiabatically on and off in the leads (from [17])

as the coupling of the spin to a fictitious in-plane magnetic field directed perpendicular to the electron momentum. Hence in a ring it points mainly in radial direction. The strength of the SO field can be tuned by an external gate voltage [25] allowing to control experimentally the spin evolution.

We assume *spin-polarized* spin-up electrons entering the ring from the right (see Fig. 1). Figure 2 displays numerically computed (see Sect. “Numerical Quantum Transport”) conductance traces as a function of the external magnetic flux  $\phi$  for weak and moderate Rashba strengths. The overall conductance is presented as a solid line, and its spin-resolved components,  $G^{\uparrow\uparrow}$  and  $G^{\downarrow\uparrow}$ , corresponding to outgoing spin-up and -down channels, are shown as dashed and dotted lines, respectively. In the weak SO coupling limit, Fig. 2a, the overall conductance (solid line) shows the usual AB oscillations of period  $\phi_0$  and is dominated by  $G^{\uparrow\uparrow}$  (dashed line). As expected for weak spin-coupling, the spin polarization is almost conserved during transport. Interesting features appear for the case of moderate SO coupling depicted in panel (b). There, both components,  $G^{\uparrow\uparrow}$  (dashed line) and  $G^{\downarrow\uparrow}$  (dotted line), contribute similarly to the overall conductance (solid line). However, the spin polarization of the transmitted electrons varies as a function of the magnetic flux  $\phi$ :  $G^{\downarrow\uparrow} = 0$  at  $\phi = 0$ , while  $G^{\uparrow\uparrow} = 0$  at  $\phi = \phi_0/2$ . Hence, for zero flux all transmitted carriers conserve their original (incoming) spin-orientation, while for  $\phi = \phi_0/2$  the transmitted particles reverse their spin polarization. In other words: by tuning the magnetic flux from 0 to  $\phi_0/2$  we can reverse the spin-polarization of transmitted particles in a controlled way. Hence the AB ring with SO coupling acts as a tunable spin-switch. This switching can be traced back to constructive or destructive AB interference [4].



**Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 2**

**Mesoscopic Aharonov-Bohm spin interference:** The conductance of spin-up polarized carriers entering a single-channel two-dimensional Aharonov-Bohm ring (see Fig. 1) is shown as a function of flux  $\phi = \pi r_0^2 B$  (in units of the flux quantum  $\phi_0 = hc/e$ ) through the ring in the presence of Rashba spin-orbit coupling. For panel a and b the scaled spin-orbit coupling strength takes the values 0.2 and 1.0, expressed as the product  $\omega_R T_0 / 2\pi$  of the precession frequency  $\omega_R$  of the spin around the effective spin-orbit magnetic field and the time  $T_0$  for traveling of the electrons around the ring. The overall conductance (solid line) is split into its components  $G^{\uparrow\uparrow}$  (dashed line) and  $G^{\downarrow\uparrow}$  (dotted line). Note in panel b the continuous change of the spin polarization, related to  $G^{\uparrow\uparrow} - G^{\downarrow\uparrow}$ , with  $\phi$  and the spin switching at  $\phi = \phi_0/2$  (adapted from [17])

Switching a given spin polarization requires the generation of spin-polarized particles in non-magnetic mesoscopic conductors in the first place. Since spin injection from ferromagnets into a semiconductor remains problematic [26], alternative proposals have been made to achieve spin-polarized currents or spin accumulation without magnets, which we will briefly review in Sect. “Spin Filtering in Nanostructures”. Among those are the spin Hall effect [27] and, in the context of coherent mesoscopic transport, concepts for Zeeman- and SO-mediated adiabatic spin pumping and spin ratchets.

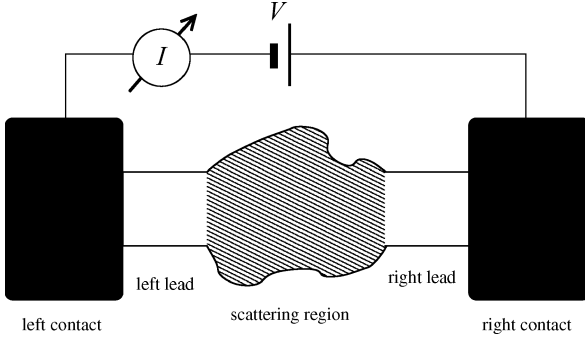
For a recent account on spin phenomena in systems of reduced dimensions see [28]; for a review on the related field of magnetization dynamics and pumping in layered magnetic heterostructures see [29].

A complete and quantitative understanding of spin phenomena in the mesoscopic realm requires computational approaches to quantum transport which also serve

as reference calculations for analytical predictions usually based on model assumptions. However, also numerical approaches cannot cope with the full many-body transport problem without relying on approximations. Here we focus on mesoscopic conductors, i. e. systems with a considerable number of electrons, with strong coupling to external leads. Then, a mean-field treatment is usually justified which allows one to reduce the Hamiltonian to a single-particle problem with an effective confinement potential resulting from a combination of external and mean-field potentials.

We further consider coherent transport close to equilibrium at relatively low bias, excluding inelastic effects, such that the Landauer approach to transport is justified. However, even in this case brute-force computational approaches quickly reach their limits: Conductors at mesoscopic scales are typically characterized by extensions which are (at least in one direction) much larger than the Fermi wavelength of the charge carriers, the shortest quantum scale involved. This implies rapidly oscillating, complex and often irregular spinor wave functions extending throughout the systems which require for the quantum mechanical numerical solution either huge sets of basis functions or, in tight-binding approaches, the use of rather fine, preferentially adapted grids. The strength of the widespread tight-binding approaches for transport discussed and reviewed here lies in their flexibility and general applicability. Moreover, tight-binding transport codes can be combined with density-functional (DFT) calculations for structure and electronic properties of the nano-conductors by using the parameters computed within DFT as input. This approach is frequently applied to transport in nano- or molecular electronics.

The paper is organized as follows: In the methodological Sect. “Numerical Quantum Transport” we will first briefly summarize and provide the key relations for spin quantum transport within the Landauer framework. In Sect. “Matrix Reordering Strategies for Quantum Transport” we focus on and explain in some detail advanced computational concepts, making use of graph-theory, to implement powerful and flexible algorithms for tight-binding transport codes. The numerical strength of the codes is demonstrated for ring-type geometries showing that by efficient tight-binding implementation one can gain orders of magnitude in performance. In Sect. “Spin Filtering in Nanostructures” and “Pure Spin Current Generation” we employ these numerical schemes to address two important aspects of spin-dependent transport, namely spin filtering and generating pure spin currents. To this end we focus on laterally-confined 2D ballistic nanostructures with Rashba SO interaction. We conclude



**Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 3**

**Schematic view of a two-point measurement setup**

with an outline of future research directions in mesoscopic spin transport which we consider important.

### Numerical Quantum Transport

#### Landauer–Büttiker Transport Theory

If the dimensions of a device get smaller than the phase coherence length  $l_\phi$  of charge carriers, classical transport theories are not valid any more. Instead, carrier dynamics is now governed by quantum mechanics and the wave-like nature of particles becomes important. In general, the conductance/resistance of such a device does not follow Ohm's law.

Consider a two-point measurement setup as shown in Fig. 3: A scattering region is connected to large (phase-breaking) reservoirs by leads. The leads are assumed to be perfect and infinitely long to define asymptotic eigenstates  $\phi_{n,\sigma}(y)e^{\pm ikx}$  at energy  $E$ , where  $n$  is the quantum number of transverse confinement —also called the *channel* number—and  $\sigma$  is the spin index. The total scattering eigenstate  $\psi_{n,\sigma}$  originating from channel  $n$  with spin  $\sigma$  in the left lead is, within the lead region, given by

$$\psi_{n,\sigma}(\mathbf{x}) = \begin{cases} \phi_{n,\sigma}(y)e^{ikx} + \sum_m r_{m,\sigma';n,\sigma} \phi_{m,\sigma'}(y)e^{-ikx} & \text{for } \mathbf{x} \text{ in left lead} \\ \sum_m t_{m,\sigma';n,\sigma} \phi_{m,\sigma'}(y)e^{ikx} & \text{for } \mathbf{x} \text{ in right lead} \end{cases}, \quad (1)$$

and obeys the stationary Schrödinger equation  $H\psi_{n,\sigma} = E\psi_{n,\sigma}$ . The conductance  $G$  in linear response can then be calculated within the *Landauer–Büttiker* formalism [30,31,32] (for tutorials see [1,33]):

$$G = \frac{e^2}{h} \sum_{n,m} \sum_{\sigma,\sigma'} T_{m,\sigma';n,\sigma} = \frac{e^2}{h} T_C, \quad (2)$$

where  $T_{m,\sigma';n,\sigma} = |t_{m,\sigma';n,\sigma}|^2$  is given by the squared transmission amplitudes of the scattering states  $\psi_{n,\sigma}$ . The fraction  $e^2/h$  is called the conductance quantum. The *scattering matrix*  $S_{m,\sigma';n,\sigma}$  is a useful definition that combines reflection and transmission amplitudes for both leads into a single matrix. In this notation the index  $n,\sigma$  then also contains information about the respective lead.

The problem of calculating the conductance  $G$  is thus reduced to calculating the scattering eigenstates  $\psi_{n,\sigma}$ . Alternatively, the scattering amplitudes  $r_{m,\sigma';n,\sigma}$  and  $t_{m,\sigma';n,\sigma}$  can also be derived from the retarded Greens function  $G^R(\mathbf{x}, \mathbf{x}')$  of the system. The retarded Greens function for a given energy  $E$  obeys the equation

$$(E - H + i\eta) G^R(\mathbf{x}, \mathbf{x}') = \delta(\mathbf{x} - \mathbf{x}'), \quad (3)$$

where  $H$  is the Hamiltonian of the system and  $\eta$  an infinitesimally small number. Formally this equation can be solved as

$$G^R = (E - H + i\eta)^{-1}. \quad (4)$$

The transmission and reflection amplitudes are then given by the *Fisher–Lee* relation [34]:

$$t_{m,\sigma';n,\sigma} = -i\hbar \sqrt{v_m v_n} \int_{C_R} dy \times \int_{C_L} dy' \phi_{m,\sigma'}(y) G^R(\mathbf{x}, \mathbf{x}') \phi_{n,\sigma}(y'), \quad (5)$$

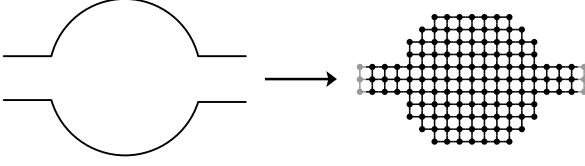
$$r_{m,\sigma';n,\sigma} = \delta_{mn} \delta_{\sigma'\sigma} - i\hbar \sqrt{v_m v_n} \int_{C_L} dy \times \int_{C_L} dy' \phi_{m,\sigma'}(y) G^R(\mathbf{x}, \mathbf{x}') \phi_{n,\sigma}(y'), \quad (6)$$

where  $v_n$  denotes the velocity of channel  $n$  and the integration runs over the cross-section  $C_L(C_R)$  of the left (right) lead. Equations (5) and (6) are valid only for leads without magnetic fields and no spin-orbit interaction. Baranger and Stone [35] have extended the formalism to also account for arbitrary magnetic fields in the leads, and their description can also be applied to finite spin-orbit interaction.

#### Tight-binding Representation of the Hamiltonian

Except for very simple geometries, the scattering problem cannot be solved analytically. Therefore, the use of computers for a numerical solution of the scattering problem is very often the method of choice. However, the related stationary Schrödinger equation  $H\psi = E\psi$  is a differential equation with continuous degrees of freedom that are difficult to treat on a computer. In general, a numerical solution is thus only attempted within a *discrete* basis set which converts the differential equation into a matrix equation.





Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 4

Discretizing a continuous region on a square grid

The method of finite differences is a very simple and yet powerful way to introduce such a discrete basis set and has been applied to the Schrödinger equation already as early as 1934 [36,37] (for an introduction see [1,33]). Here we illustrate, as an example, the application of the method for a simple one-dimensional effective mass Hamiltonian including a potential

$$H = -\frac{\hbar^2}{2m} \frac{d^2}{dx^2} + V(x). \quad (7)$$

In the method of finite differences, space is approximated by a grid of discrete lattice points spaced equidistantly with lattice constant  $a$ . For the 2D case, usually a square grid is used as depicted in Fig. 4. Using the Taylor expansion of the wave function  $\psi$  we can write

$$\begin{aligned} \psi(x+a) &= \psi(x) + \psi'(x)a + \frac{1}{2}\psi''(x)a^2 \\ &\quad + \frac{1}{6}\psi^{(3)}(x)a^3 + \frac{1}{24}\psi^{(4)}(x)a^4 + \dots, \end{aligned} \quad (8)$$

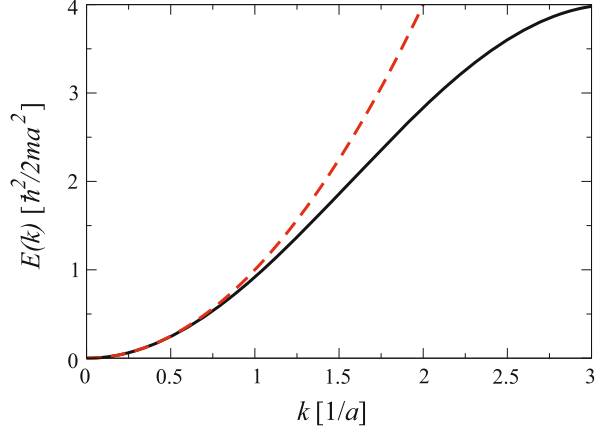
$$\begin{aligned} \psi(x-a) &= \psi(x) - \psi'(x)a + \frac{1}{2}\psi''(x)a^2 \\ &\quad - \frac{1}{6}\psi^{(3)}(x)a^3 + \frac{1}{24}\psi^{(4)}(x)a^4 + \dots. \end{aligned} \quad (9)$$

Adding Eq. (8) and (9), we arrive at an expression for the second derivative of the wave function in terms of values of the wave function on the grid,

$$\begin{aligned} \frac{d^2}{dx^2} \psi(x) &= \frac{1}{a^2} (\psi(x+a) + \psi(x-a) - 2\psi(x)) + \mathcal{O}(a^2), \end{aligned} \quad (10)$$

valid up to second order in the lattice spacing  $a$ . The differential equation

$$-\frac{\hbar^2}{2m} \frac{d^2}{dx^2} \psi(x) + V(x)\psi(x) = E\psi(x) \quad (11)$$



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 5

Energy spectrum  $E(k)$  for the continuous Schrödinger equation (red dashed line) and the tight-binding approximation (black solid line)

is thus replaced by a set of difference equations

$$-\frac{\hbar^2}{2ma^2} (\psi(x_{i+1}) + \psi(x_{i-1}) - 2\psi(x_i)) + V(x)\psi(x_i) = E\psi(x_i), \quad (12)$$

where  $x_{i\pm 1} = x_i \pm a$ , yielding the tight-binding representation of the Hamiltonian:

$$\begin{aligned} H &= \sum_{x_i} -\frac{\hbar^2}{2ma^2} (|x_i\rangle \langle x_{i+1}| + \text{h.c.}) \\ &\quad + \left( 2\frac{\hbar^2}{2ma^2} + V(x_i) \right) |x_i\rangle \langle x_i|. \end{aligned} \quad (13)$$

Here,  $|x_i\rangle$  denotes a state localized at grid point  $x_i$ .

In principle, the quality of the finite differences approximation can be improved up to a desired precision by reducing the lattice spacing  $a$ . However, since this leads to a larger problem size, the minimum lattice spacing achievable is set by the available computing time and memory. Thus, one must keep an eye on the validity of the finite differences approximation. In Fig. 5 we show the energy spectrum  $E(k)$  for the continuous one-dimensional Schrödinger equation and the tight-binding (finite differences) approximation. The tight-binding approximation only holds for  $ka \ll 1$  and  $E \ll \hbar^2/(2ma^2)$ , and it does not make sense to consider the whole energy spectrum given by the tight-binding band width. The method of finite differences presented here can straight-forwardly be applied to more complex Hamiltonians, including for example spin-orbit interactions [17], and will be later

used to calculate transport phenomena including spin in Sect. “Spin Filtering in Nanostructures” and “Pure Spin Current Generation”.

A tight-binding representation of the Hamiltonian can also be obtained by applying the finite element method [38]. Furthermore tight-binding Hamiltonians are also used in treatments beyond the effective mass approximation, such as from atomic orbitals in empirical tight-binding models [39,40,41], or from orbitals of the Kohn–Sham equations within DFT [42,43,44].

### Numerical Algorithms

Within the tight-binding approximation the Hamiltonian  $H$  can be represented by a matrix, even though this matrix is still infinite as the leads are infinitely long. However, the infinite matrix problem can be reduced to a finite problem by partitioning the system into three isolated parts: left lead, scattering region, and right lead. The Hamiltonian then reads

$$H = \begin{pmatrix} H_L & V_{LS} & 0 \\ V_{SL}^\dagger & H_S & V_{SR} \\ 0 & V_{RS} & H_R \end{pmatrix}, \quad (14)$$

where  $H_{L(R)}$  is the (infinite) Hamiltonian of the left (right) lead,  $H_S$  is the Hamiltonian of the scattering region and of finite size.  $V_{SL} = V_{LS}^\dagger$  and  $V_{SR} = V_{RS}^\dagger$  represent the coupling of the left and right lead to the scattering region. Since the leads are always chosen such that asymptotic eigenstates can be defined, the Hamiltonian of the isolated leads must contain some periodicity that facilitates calculating their Greens functions  $g_{L,R}^R$ . This can be done analytically for simple systems [1,33], for more complex situations the Greens function can be calculated numerically either by iteration [45,46] or semianalytical formulas [40,44,47]. Introducing the retarded self-energy  $\Sigma^R = \sum_{i=L,R} V_{Si} g_i^R V_{iS}$ , [1,33] the Greens function  $G_S^R$  of the scattering region is given by

$$G_S^R = (E - H - \Sigma^R)^{-1}, \quad (15)$$

reminiscent of Eq. (4) but with an effective Hamiltonian  $H + \Sigma^R$ .

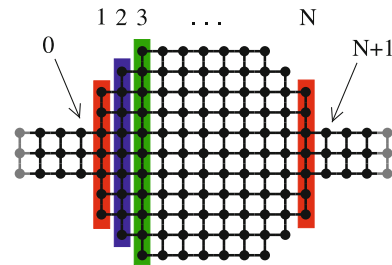
The original infinite-dimensional problem has thus been reduced to a finite size matrix problem that can, in principle, be solved straight-forwardly on a computer. However, for any but rather small problems, the computational task of the direct inversion in Eq. (15) is prohibitive. Therefore, many algorithms make use of the *sparsity* of the Hamiltonian matrix in tight-binding representation – in particular they employ the property that this matrix can

be written in block-tridiagonal form

$$H = \begin{pmatrix} \ddots & & & & & & & & & & \\ & H_L & V_L & & & & & & & & \\ & V_L^\dagger & H_L & H_{01} & & \ddots & & & & & \\ & & H_{10} & H_{11} & H_{12} & & & & 0 & & \\ & & & H_{21} & H_{22} & H_{23} & & & & \ddots & \\ & & & & H_{32} & \ddots & & & & & \\ & \ddots & & & & \ddots & & & & & \\ & & 0 & & & & H_{N-1N} & & & & \\ & & & \ddots & & & H_{NN-1} & H_{NN} & H_{NN+1} & & \\ & & & & & & & H_{N+1N} & H_R & V_R & \\ & & & & & & & & V_R^\dagger & H_R & \\ & & & & & & & & & \ddots & \end{pmatrix}. \quad (16)$$

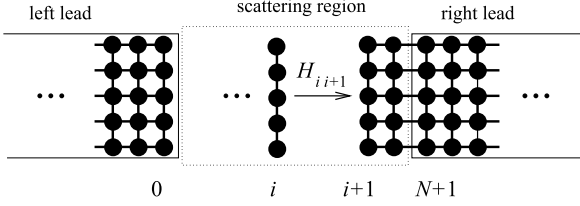
Here the index  $L(R)$  denotes the blocks in the left (right) lead,  $1 \dots N$  the blocks within the scattering region, and  $0, (N+1)$  the first block in the left (right) lead. Such a form arises naturally in the method of finite differences, when grid points are grouped into vertical slices according to their  $x$ -coordinates, as shown in Fig. 6, but also applies to any other sparse tight-binding Hamiltonians.

The block-tridiagonal form of the Hamiltonian is the foundation of several quantum transport algorithms, together with the fact that, according to Eq. (5) and (6), only the blocks  $G_{N+10}^R$  and  $G_{00}^R$  of the Greens function  $G^R$  are needed for the calculation of transmission and reflection probabilities. The transfer matrix approach applies naturally to block-tridiagonal Hamiltonians, but becomes unstable for larger systems. However, a stabilized version has been developed by Usuki et al. [48,49]. In the dec-



**Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 6**

Block-tridiagonal matrix form arising in the method of finite differences. Grid points with the same  $x$ -coordinate are placed into the same block



**Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 7**

**Schematic view of the recursive Greens function algorithm: the system is built up adding block by block**

imation technique [50,51], the Hamiltonian of the scattering region is replaced by an effective Hamiltonian between the two leads by eliminating internal degrees of freedom. The contact block reduction method [52] calculates the full Greens function of the system using a limited set of eigenstates. The recursive Greens function (RGF) technique [53,54,55] uses Dyson's equation to build up the system's Greens function block by block. It has also been adapted to Hall geometries with four terminals [56] and to calculate non-equilibrium densities [57,58]. Furthermore, the RGF algorithm has been formulated to be suitable for parallel computing [59], and the modular recursive Greens function (MRGF) method is an extension to take advantage of special geometries, such as circles or rectangles [60,61].

Here we give a simple form of the RGF algorithm as described in [33,55]. The RGF method makes use of Dyson's equation

$$G = G_0 + G_0 V G \quad (17)$$

to add successively blocks to the system, as depicted in Fig. 7. Let  $G^{R,(i)}$  denote the Greens function for the system containing all blocks  $\geq i$ . Then, at energy  $E$ , the Greens function  $G^{R,(i)}$  is related to  $G^{R,(i+1)}$  by

$$G_{ii}^{R,(i)} = (E - H_{ii} - H_{i+1,i} G_{i+1,i+1}^{R,(i+1)} H_{i,i+1})^{-1} \quad (18)$$

and

$$G_{N+1,i}^{R,(i)} = G_{N+1,i+1}^{R,(i+1)} H_{i+1,i} G_{ii}^{R,(i)}. \quad (19)$$

Starting from  $G_{N+1,N+1}^{R,(N+1)} = g_R^R$ , the surface Greens function of the right lead,  $N$  slices are added recursively, until  $G^{R,(1)}$  has been calculated. The blocks of the Greens function of the full system necessary for transport are then given by

$$G_{00}^R = \left( (g_L^R)^{-1} - H_{01} G_{11}^{R,(1)} H_{10} \right)^{-1} \quad (20)$$

and

$$G_{N+1,0}^R = G_{N+1,1}^{R,(1)} H_{10} G_{00}^R, \quad (21)$$

where  $g_L^R$  is the surface Greens function of the left lead.

Each step of the algorithm performs inversions and matrix multiplications with matrices of size  $M_i$ . Since the computational complexity of matrix inversion and multiplications scales as  $M_i^3$ , the complexity of the RGF algorithm is  $\propto \sum_{i=0}^{N+1} M_i^3$ . Thus, it scales linearly with the "length"  $N$ , and cubically with the "width"  $M_i$  of the system.

While for certain geometries the RGF algorithm cannot compete with more specialized algorithms such as MRGF, it is very versatile and easily adapted to many situations, and is thus our method of choice. In the next section we will discuss matrix reordering techniques that improve the runtime of the RGF algorithm considerably and allow the treatment of arbitrary systems.

## Matrix Reordering Strategies for Quantum Transport

### Graph-theoretical Approaches to Matrix Reordering

As shown in the previous section, the structure of a Hamiltonian matrix  $H$  does influence the runtime of the RGF algorithm. Thus, the runtime of the algorithm can potentially be improved by reordering the matrix with a permutation  $P$ ,

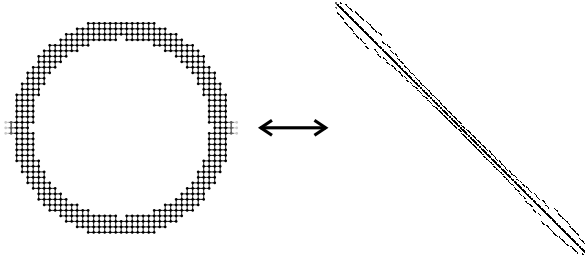
$$H' = P H P^{-1}. \quad (22)$$

At first glance, such an effort may seem pointless: For example, the block-tridiagonal structure naturally associated with a finite difference grid (as discussed in Sect. "Numerical Quantum Transport") leads to a sparse matrix with a small bandwidth, as shown in Fig. 8. However, as we show below, the choice of a suitable permutation  $P$  can still lead to a significant speed-up of the RGF algorithm.

For this we define a weight  $w(H)$  associated with a given matrix  $H$  as

$$w(H) = \sum_{i=0}^{N+1} M_i^3 \quad \text{where } M_i \text{ is the size of block } H_{ii}. \quad (23)$$

Optimizing the matrix for the RGF algorithm is then equivalent to minimizing the weight  $w(H)$ . Since  $\sum_{i=0}^{N+1} M_i = N_{\text{grid}}$ , where  $N_{\text{grid}}$  is the total number of grid points,  $w(H)$  is minimal, if all  $M_i$  are equal, and  $M_i = N_{\text{grid}}/(N+2)$ . Therefore, a matrix tends to have



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 8

Example of a sparsity structure of a matrix associated with a finite difference grid. Black dots mark non-zero entries. The picture of the finite difference grid can also be interpreted as a graphical representation of a graph

small weight, if the number  $N$  of blocks is large, and all blocks are equally sized. The reordering problem of the matrix  $H$  is thus summarized as follows:

**Matrix reordering problem:** Find a reordered matrix  $H'$  such, that

1.  $H'_{00}$  and  $H'_{N+1N+1}$  are blocks given by the left and right leads (as required by the RGF algorithm),
2.  $H'$  is block-tridiagonal ( $H'_{ij} \neq 0$ , iff  $j = i + 1, i, i - 1$ ),
3. the number  $N$  of blocks is as large as possible, and all blocks are equally sized.

These requirements define the optimization problem of reordering the matrix  $H$ . Usually, in such optimization problems finding the best solution deterministically is prohibitively expensive, and one has to resort to heuristic strategies.

In order to do that, we reformulate our matrix problem using concepts from graph theory. A *graph*  $G$  is a pair  $G = (V, E)$ , where  $V$  is a set of *vertices*  $i$ , and  $E$  a set of pairs of vertices  $(i, j) \in V \times V$ . Such a pair is called an *edge*. A graph is called *undirected*, if for every edge  $(i, j) \in E$  also  $(j, i) \in E$ . Two vertices  $i$  and  $j$  are called *adjacent*, if  $(i, j) \in E$ . A graph can be visualized by drawing dots for each vertex  $i$  and lines connecting these dots for every edge  $(i, j)$ . It should be noted, that all the pictures of finite difference grids shown so far can directly be interpreted as graphs. There is a natural one-to-one correspondence between graphs and the structure of sparse matrices. For a given  $n \times n$  matrix  $H$ , we define a graph  $G = (V, E)$  with  $V = \{1, \dots, n\}$  and  $(i, j) \in E$  iff  $H_{ij} \neq 0$ . Thus, the symmetric zero-nonzero structure of Hermitian matrices, as considered in quantum transport, leads to associated undirected graphs. An example of such a correspondence between a graph and a matrix has already been shown in Fig. 8. With respect to matrices, graphs are also very con-

venient for storing and handling sparse matrix data structures on a computer.

In terms of graph theory, matrix reordering corresponds to renumbering the vertices of a graph. Since we are only interested in reordering the matrix in terms of matrix blocks (the order within a block should not matter too much), we define a *partitioning* of  $G$  as a set  $\{V_i\}$  of disjoint subsets  $V_i \subset V$  such that  $\bigcup_i V_i = V$  and  $V_i \cap V_j = \emptyset$  for  $i \neq j$ . Using these concepts, we can now reformulate the original matrix reordering problem into a graph partitioning problem:

**Graph partitioning problem:** Find a partitioning  $\{V_0, \dots, V_{N+1}\}$  of  $G$  such that:

1.  $V_0$  and  $V_{N+1}$  contain the vertices belonging to left and right leads,
2. there are edges between  $V_i$  and  $V_j$  iff  $j = i + 1, i, i - 1$  (block-tridiagonality),
3. the number of sets  $V_i$  is as large as possible and all sets  $V_i$  have the same cardinality. Such a partitioning with all  $V_i$  equally sized is called *balanced*.

A partitioning that meets requirement 2 is called a *level set* with levels  $V_i$  and appears commonly as an intermediate step in algorithms for bandwidth reduction of a matrix [62,63,64,65]. These algorithms seek to find a level set of minimal width, i. e.  $\max_{i=0 \dots N+1} |V_i|$  as small as possible which is equivalent to requirement 3. The major difference between our graph partitioning problem and the bandwidth reduction algorithms is requirement 1: In our case the start and end blocks are given by the geometry of the system, whereas in the bandwidth reduction methods these can be chosen freely. The implications of this difference will be discussed below.

The bandwidth reduction algorithms start with the observation that a *breadth first search* (BFS) starting from any vertex in the graph creates a level set: In our situation, the BFS starts from level  $V_0$  of the left lead. Then, successively for  $i = 0, 1, 2, \dots$ , all vertices adjacent to  $V_i$  that have not been assigned to a level yet are placed in  $V_{i+1}$ . This construction ensures that each level  $V_i$  only has edges connecting to vertices in  $V_{i+1}, V_{i-1}$  and thus ensures block-tridiagonality. The search stops, once a vertex adjacent to the right lead is encountered, and all unassigned vertices are placed into the last level  $V_N$ . The number of BFS steps determines therefore the maximum number  $N$  of blocks and is related to the minimum distance between left and right lead in the graph. However, this construction of the level set also suffers from a serious problem: Depending on the distance between the leads, the last level  $V_N$  can potentially contain a large number of vertices leading to a very unbalanced partitioning. In the bandwidth reduc-

tion methods, the first and the last vertex are chosen to have (to a good approximation) maximum distance, and thus this problem does not occur there. Hence, conventional bandwidth reduction algorithms can only be applied to quantum transport problems, if the leads are—in terms of the underlying graph—furthest apart.

In this study, we will consider two reordering algorithms: First, the Gibbs–Poole–Stockmeyer (GPS) algorithm [65], a widely used bandwidth reduction algorithm. The GPS algorithm combines the level sets originating from a BFS from both left and right lead to give an optimized level set. Due to the limitations discussed above, it can only be used efficiently for systems with leads far apart. To overcome this difficulty partly, we also propose a second algorithm, later referred to as *BFS partitioning*: The system is bisected recursively by means of a simultaneous BFS from left and right leads. In a bisection process, vertices that are closer (as given by the number of BFS steps) to the left (right) lead are placed into the left (right) level. The resulting two levels are then further bisected recursively until the final level set has been constructed. This algorithm tries to avoid an unbalanced partitioning, as every step tries to create a balanced bisection. We have found this global approach—as opposed to the local approach in the BFS—to yield balanced partitionings for systems where there are only few local minima in the weight  $w(H)$ . For general systems, a more sophisticated method should be used [66].

Obviously, reordering the matrix will only improve the runtime of the RGF algorithm, if the overhead of the reordering process is small compared to the actual transport calculation. Because of this reason, applying general optimization algorithms to the matrix reordering problem is not an option. Instead, heuristics designed for graph problems give much better performance. The GPS algorithm scales linearly with the number of edges  $|E|$ , and since in a tight-binding representation  $|E| \propto N_{\text{grid}}$ , its computational complexity is  $\mathcal{O}(N_{\text{grid}})$ , whereas the BFS partitioning algorithm scales as  $\mathcal{O}(N_{\text{grid}} \log N_{\text{grid}})$ . In any case, the scaling is much more favorable than that of the RGF algorithm,  $\propto \sum_{i=0}^{N+1} M_i^3$ , so that for systems of typical size the overhead of the reordering process becomes negligible, as we will demonstrate in the next section.

### Example: Ring Geometry

In order to demonstrate the performance of the algorithms discussed above, we consider their application to a ring geometry in finite difference approximation. The performance of the RGF algorithm after matrix reordering is compared with the performance using the ordering that

arises naturally in finite difference grids as shown in Fig. 6 (in the remainder referred to as *natural partitioning*).

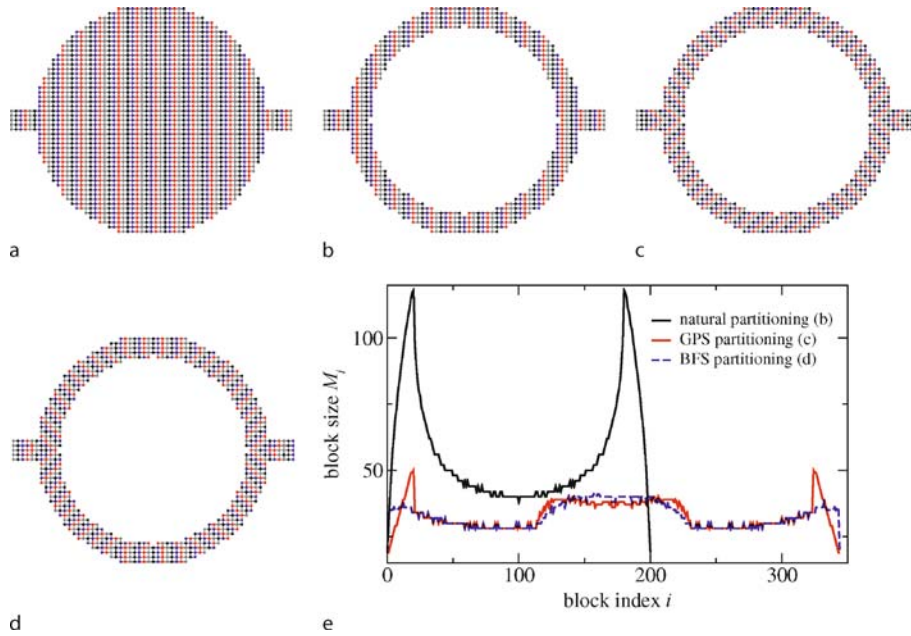
In Fig. 9a–d we show four different approaches for calculations in a ring geometry. A ring can be treated as a circular cavity (see Fig. 9a), with a large potential on the lattice points inside the inner ring diameter. This approach is easier to implement than a real ring but less efficient, as a large additional number of lattice points enters the calculation. However, this approach has been used frequently in the past, and therefore we also consider its performance. Transport calculations in a real ring require somewhat more bookkeeping because of the non-trivial geometry, but can be easily done describing the grid as a graph. For the circular cavity, we only consider natural partitioning, as shown in Fig. 9a, for the ring we consider natural partitioning (Fig. 9b), GPS partitioning (Fig. 9c) and BFS partitioning (Fig. 9d).

The partitionings in Fig. 9c and d are dramatically different from the natural partitioning. The levels align mainly in vertical or diagonal directions as these are the preferred directions in the square lattice. The number of levels is increased with respect to the natural partitioning, as the distance between both leads is much larger than the number of vertical slices, leading to a larger number of blocks in the block-tridiagonal matrix, as can be seen from Fig. 9e. Both GPS and BFS partitioning lead to a drastically reduced block size with respect to the natural partitioning and the result is rather balanced. Though the actual partitionings in Fig. 9c and d look quite different, with respect to minimizing the weight  $w(H)$  they are equally good. The BFS partitioner conforms to the geometric structure, as it puts vertices in levels according to their distance from the leads. Except for a small number of blocks in the beginning and the end of the block-tridiagonal matrix (see Fig. 9e), the GPS partitioner leads to an equally balanced structure, although the partitioning looks quite different. The GPS partitioner works well in this case, as the leads are almost at maximum distance in this ring structure.

We now apply the recursive Greens function algorithm from Sect. “Numerical Quantum Transport” to the different partitionings. In Fig. 10a we show the runtime of the algorithm as a function of the number of lattice points in the leads, which is also the number of lattice points across the arms of the ring. Note that in all cases the runtime includes both the time spent in calculating the matrix reordering and the time spent in the actual transport calculation.

The runtime scales similarly in all cases, as this is the scaling of the RGF algorithm. Nevertheless, the runtimes of the different approaches can differ by a factor that is significant. As expected, the circular cavity is slowest, due





Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 9

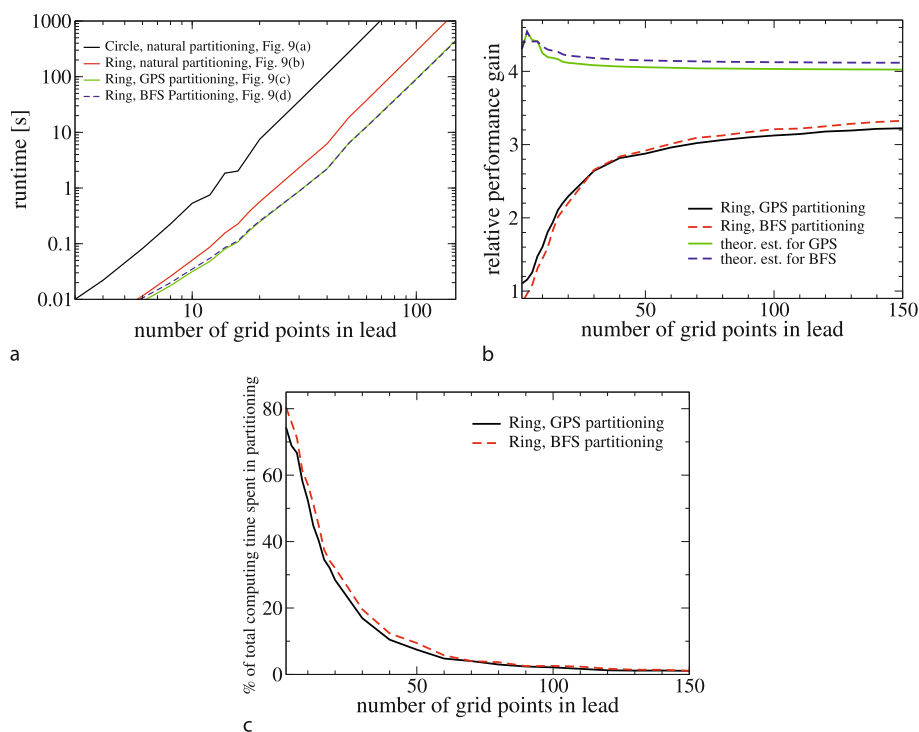
Partitioning of the different systems considered in the performance study. **a** Circular cavity in natural partitioning, where the ring structure is enforced by a potential term in the middle of the cavity, **b** ring in natural partitioning, **c** ring in GPS partitioning and **d** ring in BFS partitioning. In **a**–**d**, vertices belonging to the same level in the partitioning are marked with the *same color*, different levels are marked in *alternating colors*. Note that, in order to reveal the partitioning structure more clearly, the grids shown here are much smaller than in a typical calculation. **e** Size  $M_i$  of the blocks  $H_{ij}$  in the block-tridiagonal matrix for a ring with 20 lattice points in the leads for natural, GPS and BFS partitioning (the circular cavity is not shown here)

to the extra number of lattice points. GPS and BFS partitionings lead to a rather similar performance that is *significantly* better than the ring in natural partitioning. It outperforms the circular cavity even by a factor of up to 100. In the remainder, we examine the performances of the ring for different partitionings in more detail and leave out the circular cavity. In Fig. 10b we show the relative performance gain of GPS and BFS partitionings over the natural partitioning. Except for the smallest of systems that are too small to be useful in practice, the performance of GPS and BFS partitionings is better than the natural partitionings. Even for moderately sized systems the performance gain through the matrix reordering is approximately 3, with the BFS partitioner being slightly better than the GPS partitioner. In Fig. 10b we also show estimates of the performance gain calculated from the weights  $w(H)$  of the different partitionings. These estimates predict a performance gain of approximately 4. For small system sizes, we see deviations from these estimates because of the overhead of the partitioning process, for larger system sizes we almost reach the estimated value. On modern computer architectures, runtime does not only depend on the number of arithmetic operations [67], and thus we do not achieve the full theoretical potential of the reordering, yet still sig-

nificant improvements. In Fig. 10c we show the fraction of time spent in calculating the matrix reordering with respect to the total computation time, and as expected the overhead becomes negligible already for moderately sized systems. It should be noted that in actual calculations the partitioning is commonly only done once, and transport calculations can be done repeatedly with the same partitioning: Usually one is interested in transport properties depending on some parameters, and these generally do not change the *structure* of the Hamiltonian matrix but only the values of the respective entries. In this case, the partitioning overhead becomes even more irrelevant.

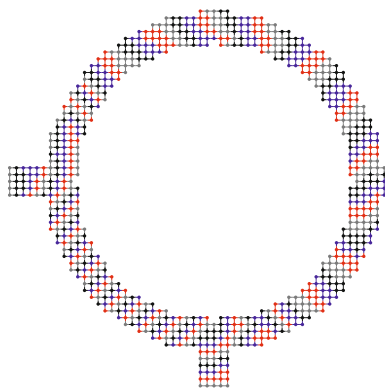
It should be emphasized, that for all the situations, the *same* transport code was used. In addition to the significant speedup through the graph techniques considered here, the abstraction of the system through graph structures allows for very generic transport codes. This is an additional strength of this approach, as the well-established RGF algorithm is thus readily applied to arbitrary systems that would require special treatment otherwise, such as a scattering region with perpendicular leads, as depicted in Fig. 11.

Of course, for any system, an algorithm designed for that special system will probably outperform generic ap-



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 10

Performance of the partitionings (for rings, Fig. 9): **a** runtime of the different systems as a function of the number of lattice points in the leads, i. e. as a function of system size. The calculations were performed on a Core2Duo T5500 processor and 1GB of memory, and the runtime includes both partitioning overhead and transport calculation. **b** Performance gain of GPS and BFS partitionings with respect to natural partitioning and **c** overhead of the matrix reordering as a function of system size



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 11

BFS partitioning of a ring with perpendicular leads. Note that in this case the two leads are closer together and thus the number of blocks is reduced. Therefore the block sizes tend to be larger than in the previous examples

proaches. However, the combination of matrix reordering and RGF algorithm can be applied to arbitrary systems and is thus probably the most versatile transport approach. In addition, all the algorithms relying on the block-tridiagonal

structure of the Hamiltonian mentioned in Sect. “Numerical Quantum Transport” can benefit from these matrix reordering strategies.

Modern transport calculations tend to be more and more complex and time-consuming. For example, an electronic calculation including the electron spin typically takes  $2^3 = 8$  times longer than a calculation on spinless electrons. Furthermore, calculations including disorder involve averages over many disorder configurations. Any increase in computation speed is therefore beneficial, and the added versatility through matrix reordering methods makes these techniques even more useful.

### Spin Filtering in Nanostructures

In the introductory example in Sect. “Introduction” it was shown how to realize systems that work as spin switches making use of the interference of wavefunctions propagating clockwise and counterclockwise in Aharonov–Bohm rings with SO-interaction. However, several other device proposals have been put forward utilizing different concepts in order to achieve spin filtering in mesoscopic systems.

A very prominent category is transverse focusing of ballistic electrons/holes in two dimensional electron/hole gases (2DEGs/2DHGs) [68]. In materials that exhibit SO interaction the cyclotron radius of ballistic electrons/holes due to a magnetic field perpendicular to the 2DEG/2DHG depends on the spin state of the charge carriers. Therefore, it is possible to filter out either spin-up or spin-down electrons by an appropriate arrangement of quantum point contacts and a proper choice of the perpendicular magnetic field. Apart from several theoretical treatments of this topic [69,70,71], spin filtering by transverse focusing has already been experimentally verified in a GaAs-based 2DHG [72]. In related experiments, spin-polarized currents in 2DEGs/2DHGs were detected by a setup consisting of point contacts and making use of transverse focusing of the charge carriers [73]. With such a detector it was possible to confirm the presence of spin-polarized currents emitted from mesoscopic quantum dots utilizing quantum interference at high in-plane magnetic fields [74] and from quantum point contacts, which were either made spin sensitive with high in-plane magnetic fields [73] or showed a pronounced “0.7-anomaly” [75].

A further appealing approach to filter spins are three terminal structures, that act as mesoscopic Stern–Gerlach type spin filters [76]. In these devices one of the leads injects spin-unpolarized current and, after passing a region where the spin-degeneracy is lifted, oppositely-polarized output currents exit through the other two leads. This separation of up and down spins can be accomplished, e.g., by utilizing Rashba SO interaction [77,78,79].

However, three terminals are not required to create spin polarized currents. Many devices, as e.g. the AB-ring presented in Sect. “Introduction”, typically rely on two terminals only, where transport through tailored geometries with SO interaction [80,81,82], magnetic fields [83,84,85,86] or a combination of both [87] can result in a significant spin filter effect.

As a representative example for the methods mentioned above, in the present section we present spin filtering due to Rashba SO interaction in quantum wires connected to two terminals. We consider a quantum wire in  $y$ -direction realized in a 2DEG in the  $(x, y)$  plane connected to two nonmagnetic leads. The Hamiltonian of the system, with spatially dependent Rashba SO interaction is given by

$$H_0 = \frac{\hat{p}^2}{2m^*} + \frac{\alpha(x)}{2\hbar}(\hat{\sigma}_x \hat{p}_y - \hat{\sigma}_y \hat{p}_x) + (\hat{\sigma}_x \hat{p}_y - \hat{\sigma}_y \hat{p}_x) \frac{\alpha(x)}{2\hbar} + V(x, y) + U_B(x, y). \quad (24)$$

Here  $V(x, y)$  is the lateral transverse confinement potential forming the quantum wire, while  $U_B(x, y)$  is an additional electrostatic potential in the system, e.g., realized by gate voltages. Furthermore,  $\hat{\sigma}_i$  denote the Pauli spin matrices, and  $m^*$  is the effective electron mass of the semiconducting material. We consider a constant Rashba SO interaction strength ( $= \alpha_C$ ) in the central region of the system, which is connected to two semi-infinite leads on opposite sides, where  $\alpha(x)$  is chosen to be zero avoiding ambiguities in the definition of spin current that arise for leads with SO interaction [88]. In order not to introduce additional effects due to an abrupt jump in the SO coupling strength, the parameter  $\alpha(x)$  is changed sufficiently smooth from zero to  $\alpha_C$  between the leads and the central region. For the numerical calculations presented in the next two sections the Hamiltonian (24) is discretized as shown in Sect. “Numerical Quantum Transport” yielding a tight-binding Hamiltonian on a square grid.

In the rest of this section we investigate the transport properties in the linear response regime due to an infinitesimal bias voltage  $\delta U$  applied between the left and right contact. The charge (C) and spin (S) current in the Landauer–Büttiker formalism are then given by

$$I_{C/S} = G_{C/S} T_{C/S} \delta U,$$

where  $G_C = e^2/h$  and  $G_S = e/4\pi$  are the conductance quanta of charge and spin respectively. Since, opposite to charge current, the spin current can be different in the right and left lead [85], here we choose to evaluate the respective currents in the right lead. Then the transmission probabilities  $T_{C/S}$  at the Fermi energy are given by

$$T_C = T_{+,+} + T_{+,-} + T_{-,+} + T_{-,-}, \quad (25)$$

$$T_S = T_{+,+} + T_{+,-} - T_{-,+} - T_{-,-},$$

where  $T_{\sigma,\sigma'} = \sum_{n \in R, n' \in L} |S_{n,\sigma;n',\sigma'}|^2$  is the probability for an electron, injected into the left (L) lead with spin state  $\sigma'$  to be transmitted to the right (R) lead and end up there in spin state  $\sigma$ . In the present and the following section we fix the spin quantization axis to the  $y$ -axis. The scattering matrix elements  $S_{n,\sigma;n',\sigma'}$  and therefore also the spin resolved transmission probabilities  $T_{\sigma,\sigma'}$  are evaluated using the recursive Greens function algorithm presented in Sect. “Numerical Quantum Transport” and “Matrix Reordering Strategies for Quantum Transport”.

One general feature of Landauer transport in a quantum wire with SO interaction and non-magnetic leads is the absence of spin-polarized currents in a lead that supports only a single transversal mode. This property can be derived from the invariance of the system under the time-

reversal operator  $\hat{T} = -i\hat{C}\sigma_y$  [89], where  $\hat{C}$  is the operator of complex conjugation. For a perfect quantum wire which is translationally invariant in the direction of transport all occupied transversal subbands transmit without reflection and spin polarization is not possible due to SO interaction. However, if backreflection, caused by deviations from a perfect quantum wire, is present, it is possible to observe spin-polarized currents in leads with at least two transversal channels. There the typical mechanism responsible for spin polarization is the mixing of spins from different transversal subbands due to the SO interaction.

In Eq. (24) this translational invariance in  $y$ -direction is already broken by the spatially varying SO interaction  $\alpha(x)$  even if the quantum wire was perfect in  $y$ -direction otherwise, i. e.  $V(x, y) = V(y)$  and  $U_B(x, y) = U_B(y)$ . However, if the region where  $\alpha(x)$  is turned on/off is sufficiently long, reflection due to the change of  $\alpha(x)$  is negligible.

There exist several device proposals relying on this mixing of spins from different transversal subbands due to  $x$ -dependent lateral confinement potentials  $V(x, y)$  or other superimposed electrostatic potentials  $U_B(x, y)$ . These device designs include, e. g., constrictions [80,90], lateral side pockets [81], or electrostatic barriers [82], to name only a few.

In most of those proposals, systems symmetric with respect to inversion of the  $x$ -coordinate were considered, i. e.  $V(x, y) = V(-x, y)$ ,  $U_B(x, y) = U_B(-x, y)$  and  $\alpha(x) = \alpha(-x)$ . Then the Hamiltonian (24) is left invariant upon application of the symmetry operation

$$\hat{P} = -i\hat{C}\hat{R}_x\hat{\sigma}_z, \quad (26)$$

where the operator  $\hat{R}_x$  inverts the  $x$ -coordinate. The scattering wavefunctions inside the leads are changed by the operator  $\hat{P}$  in the following way:  $\hat{R}_x$  exchanges the leads, i. e., a transversal mode index  $n$  is replaced by the corresponding mode index  $\tilde{n}$  in the other lead. The operator of complex conjugation transforms incoming (outgoing) states into outgoing (incoming) states with complex conjugated amplitude. Moreover, the combined effect of  $\hat{C}\hat{\sigma}_z$  is to rotate a spinor with the coordinates  $(\theta, \phi)$  on the Bloch sphere to the coordinates  $(\theta, -\phi + \pi)$ . Exploiting this invariance of the Hamiltonian one can derive the relation  $S_{n,\sigma;\tilde{n}',\sigma'}(E) = S_{\tilde{n}',\sigma';n,\sigma}(E)$  between the elements of the scattering matrix (see [85,89] for related expressions). This results in the equality of the spin flip transmissions  $T_{+,-} = T_{-,+}$ . Therefore, for those devices to be able to work as a spin filter, in view of Eq. (25) the spin conserving transmissions  $T_{+,+}$  and  $T_{-,-}$  need to be different.

As an example, in the following we consider Landauer transport through a Rashba SO quantum wire with a con-

striction. Similar calculations were carried out in [71,80] where it was shown, that this setup is able to produce a spin polarized current of sizeable quantity. It was conjectured, that the mechanism responsible for the spin polarization was the depletion of higher transversal modes of the wire and a subsequent spin dependent repopulation of those modes when traversing the constriction [80]. To experimentally observe the predicted spin polarization the use of a transverse electron focusing technique was suggested [71].

A typical grid (with lattice spacing  $a$ ) used in the calculation for the symmetric point contact in a wire of width  $W = 15a$  is shown in Fig. 12a, where the constriction of length  $L_{PC} = 10a$  is formed by hard-wall potentials:

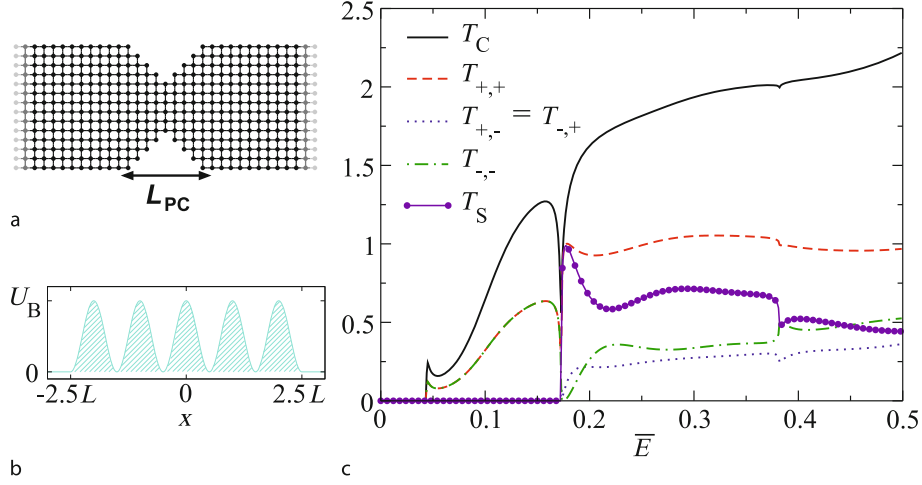
$$V(x, y) = \begin{cases} 0 & \text{for } C(x) < y < 15a - C(x) \\ \infty & \text{else} \end{cases}$$

with

$$C(x) = \begin{cases} 2.05a \left( 1 - \cos \left( \frac{2\pi \left( x + \frac{L_{PC}}{2} \right)}{L_{PC}} \right) \right) & \text{for } |x| < \frac{L_{PC}}{2} \\ 0 & \text{otherwise} . \end{cases} \quad (27)$$

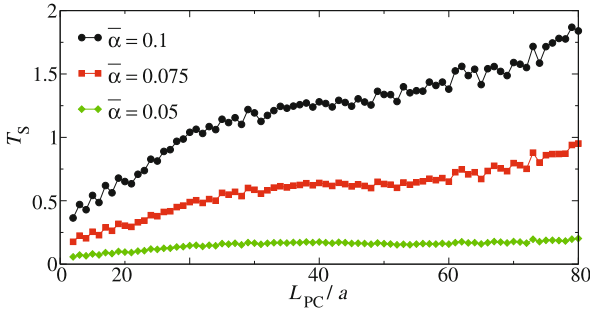
Additionally  $U_B(x, y)$  is set to zero. In Fig. 12c we present the relevant transmission probabilities with respect to the Fermi energy  $E$  for this system. There we observe that the total conductance  $T_C$  is reduced in comparison with that of a perfect quantum wire. In the latter case  $T_C$  exhibits sharp steps due to conductance quantization [91,92] which are washed out here due to tunneling processes through the constriction. Furthermore, for energies, where only a single transversal mode is supported in the quantum wire, the spin transmission vanishes as expected, i. e.  $T_{+,+} = T_{-,-}$  for energies  $\bar{E} \lesssim 0.175$ . Also the relation  $T_{+,-} = T_{-,+}$  is fulfilled as required by the symmetry of the setup. Finally, at energies where a new transversal mode opens up ( $\bar{E}_2 \approx 0.18$ ,  $\bar{E}_3 \approx 0.39$ ) dips in the transmission probabilities become apparent. Those dips can be explained by interference of localized states in the central region where  $\alpha(x) = \alpha_C$  and the scattering states in the quantum wire where  $\alpha(x) = 0$  [93].

In order to study the influence of the adiabaticity of the constriction on the degree of spin polarization that can be extracted, in Fig. 13 we show  $T_S$  as a function of the length of the constriction  $L_{PC}$  for several values of  $\alpha_C$ . For all of the curves we clearly observe an increase in spin transmission with increasing  $L_{PC}$ , in accordance with the mechanism suggested in [80]. There it was argued in the context of Landau-Zener transitions that the repopulation of higher transversal subbands will be more efficient for more



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 12

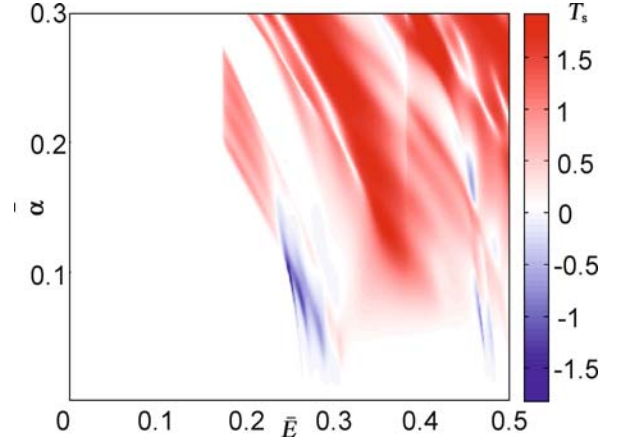
Panel a The square lattice discretization of a quantum wire (width  $W = 15a$ ) with a single constriction of length  $L_{PC} = 10a$ , see also Eq. (27). Panel b Periodic array of  $N = 5$  electrostatic barriers with period length  $L$  and barrier height  $U_B$ . Panel c Charge  $T_C$ , spin  $T_S$  and spin resolved transmission probabilities  $T_{\sigma,\sigma'}$  for the system depicted in panel a) and specified in the text at fixed SO interaction strength  $\tilde{\alpha} = [(m^*a)/\hbar^2] \alpha_C = 0.1$  with respect to the Fermi energy  $\bar{E} = [(2m^*a^2)/\hbar^2] E$ . The  $n$ th transversal mode in the wire opens at  $\bar{E}_n = (\pi^2 n^2)/(W/a)^2$



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 13

Spin transmission probability  $T_S$  at fixed injection energy  $\bar{E} = 0.25$  within the second transversal mode for three different SO interaction strengths  $\tilde{\alpha} = 0.1$  (black dots),  $\tilde{\alpha} = 0.075$  (red squares) and  $\tilde{\alpha} = 0.05$  (green diamonds) with respect to the length of the constriction  $L_{PC}$

adiabatic constrictions or barriers, resulting in a higher degree of spin polarization. For  $\tilde{\alpha} = [(m^*a)/\hbar^2] \alpha_C = 0.1$  the spin transmission even approaches the highest possible value  $T_S = 2$ . One drawback of the presented system is its restriction to unidirectional spin polarization. In agreement with the model of [80], Fig. 12 and 13 give evidence that  $T_S \geq 0$  for the parameter range considered. This limitation to output current with fixed spin polarization direction restricts the usability of the spin filter to special purposes. A possible way to circumvent this constraint is the use of a periodic array of electrostatic barriers [82], which



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 14

Spin transmission probability  $T_S$  plotted as a function of the Fermi energy  $\bar{E}$  and SO interaction strength  $\tilde{\alpha}$  for a quantum wire with  $N = 5$  electrostatic barriers of length  $L = 10a$  and height  $\bar{U}_{Barr} = [(2m^*a^2)/\hbar^2] U_{Barr} = 0.2$

we now briefly investigate. Figure 14 shows the spin transmission of a straight hard-wall quantum wire ( $C(x) = 0$  in Eq. (27)) subject to  $N = 5$  electrostatic barriers,

$$U_B(x, y) = U_B(x) = \begin{cases} \frac{U_{Barr}}{2} \left( 1 - \cos \left( \frac{2\pi(x + \frac{L}{2})}{L} \right) \right) & \text{for } -N\frac{L}{2} < x < N\frac{L}{2} \\ 0 & \text{otherwise,} \end{cases}$$



(as shown in Fig. 12b). The spin transmission is plotted as a function of the Fermi energy and the SO interaction strength. Again, one observes  $T_S = 0$  for energies below  $\bar{E}_2$ . Furthermore, different regions in parameter space exhibit opposite sign of  $T_S$ , enabling to change the sign of the output polarization, e.g., by tuning  $\alpha_c$  via gate voltages [25]. This additional functionality is due to resonant tunneling, which is absent for a quantum wire with only a single constriction or barrier.

### Pure Spin Current Generation

In the preceding sections we focused on mesoscopic geometries that exhibit functionalities such as spin filtering or spin switching when applying an external bias between the contacts in the system. In the case of a static dc bias in a two terminal geometry those spin currents arise due to different currents of spin-up and spin-down electrons flowing into the direction of the contact with the lower chemical potential. However, over the last few years a growing number of device proposals has been put forward, that exhibit the interesting feature of pure spin current generation, i.e. spin currents in the absence of net charge transport. This intriguing case appears, when the direction of motion of spin-up electrons is opposite to the direction of motion of spin-down electrons and both currents equal in absolute magnitude.

Among the devices sharing the prospect of creating pure mesoscopic spin currents are systems with more than two terminals realized in 2DEGs exhibiting Rashba [2,3] and/or the Dresselhaus [94] SO coupling. Here a charge current is induced in this multiterminal structure by the application of bias voltages between the different contacts of the system. However, if the voltage of one of the leads is adjusted to make it work as a voltage probe, no charge current passes this lead but a pure spin current can appear owing to the SO coupling present in the system. The basic working principle behind these devices is the so-called mesoscopic spin Hall effect [95], a version of the intrinsic spin Hall effect [27], where the typical system size does not exceed the phase coherence length of the electrons. The SO interaction leads to different transport dynamics for different spin species, which can be used to extract the desired pure spin currents by a clever design of the multiterminal geometry [21,96,97].

Complementary to the generation of pure spin current in multiterminal geometries, there are other types of devices not relying on the application of a net dc-bias. In spin pumping, the cyclic variation of two or more system parameters, such as e.g. gate voltages, induces spin-polarized currents at zero bias, where the induced charge cur-

rent can be tuned to disappear, leaving pure spin currents. Several realizations of spin pumps in mesoscopic systems have been proposed, relying on SO interaction [98,99] or the Zeeman coupling of electrons to external magnetic fields [100]. The latter proposal has been experimentally confirmed [101] by detecting spin-polarized currents making use of a transverse electron focusing technique [73] mentioned in the previous section.

Complementary to pumps, ratchets only require a single driving parameter to achieve directed transport, and the current direction can be switched upon tuning external parameters such as temperature. In addition to the requirement of a broken spatial symmetry the ratchet has to be operated out of thermal equilibrium. The concept of particle ratchets, which has been addressed in numerous works [102], has recently been extended to systems called spin ratchets. To be specific, the mesoscopic spin ratchets proposed so far [82,85,103], are based on a quantum wire realized in a 2DEG. Between the two contacts attached to the quantum wire an ac bias voltage  $U_R(t)$  is applied (rocking ratchet) with zero net (time-averaged) bias. Furthermore, in the central region of the quantum wire the spin degeneracy is lifted due to either SO interaction [82] or the Zeeman coupling to an external magnetic field [85,103]. Upon appropriate choice of the system geometry and tuning of the external driving the charge transported in the forward ( $U_R > 0$ ) and backward bias ( $U_R < 0$ ) situation can be made equal allowing for spin currents in the absence of net charge transport.

In the following we outline the model for the spin ratchets introduced in [82,85,103]. There, driving with a period  $t_0$  is considered. It is implied, that this period is much larger than characteristic time scales related to the electron transport through the quantum wire. Therefore, the system is assumed to be in a steady state at every instance of time, and the Landauer–Büttiker approach to transport is used for the computation of the ratchet currents. To be specific, we consider an unbiased square wave driving  $U_R(t) = U_0 \text{sign}[\sin(2\pi t/t_0)]$ , where  $U_R(t)$  is restricted to the values  $\pm U_0$ . The net current is then given by the average of the steady-state currents in the two rocking situations (labeled  $+U_0$  and  $-U_0$  respectively) for both charge and spin,

$$\langle I_{C/S}(U_0) \rangle = \frac{I_{C/S}(+U_0) + I_{C/S}(-U_0)}{2}. \quad (28)$$

Since the spin ratchet effect requires nonlinear transport [103], i.e. finite bias voltages, the Hamiltonian (24) introduced in the previous section has to be extended to additionally include the effective electrostatic potential in the conductor due to the applied bias. Therefore, we add

the term  $H_R = eU_R g(x, y; U_R)$  to the Hamiltonian (24) yielding the full Hamiltonian at finite bias:

$$H = H_0 + H_R. \quad (29)$$

Furthermore at finite bias a generalized version of the expressions for charge and spin current valid at  $U_R \neq 0$  has to be used. For coherent Landauer transport those currents can be obtained from an integration of the transmission probabilities over the Fermi window [1]. Finally, the averaged charge  $\langle I_C \rangle$  and spin  $\langle I_S \rangle$  currents can be written as [85]

$$\langle I_{C/S}(U_0) \rangle = \frac{G_{C/S}}{2e} \int_{E_C}^{\infty} dE \Delta f(E, U_0) \Delta T_{C/S}(E, U_0).$$

Here  $E_C$  is the energy of the conduction band edge and  $\Delta f(E, U_0) = [f(E, E_F + eU_0/2) - f(E, E_F - eU_0/2)]$  is the difference between the Fermi functions of the leads at bias voltage  $U_0$ , defining the Fermi window. The averaged charge/spin transmission is just the difference between the steady state transmissions in the two rocking situations:

$$\Delta T_{C/S}(E, U_0) = T_{C/S}(E, +U_0) - T_{C/S}(E, -U_0). \quad (30)$$

Considering the Hamiltonian (29) we now show under what conditions the net charge transported after one full rocking period is zero, i.e.,  $\langle I_C(U_0) \rangle = 0$ . If the electrostatic potentials  $V(x, y)$ ,  $U_B(x, y)$  and the Rashba SO strength  $\alpha(x)$  are invariant under inversion of the  $x$ -coordinate,

$$\begin{aligned} V(x, y) &= V(-x, y), \\ U_B(x, y) &= U_B(-x, y), \\ \alpha(x) &= \alpha(-x), \end{aligned} \quad (31)$$

it is appropriate to assume that the electrostatic potential distribution due to the finite applied voltage  $g(x, y; U_R)$  also possesses this symmetry. Then the total Hamiltonian (29) is invariant under the action of the symmetry operation  $\hat{P} = -i\hat{C}\hat{R}_U\hat{R}_x\sigma_z$  where  $R_U$  switches the sign of the bias voltage ( $\pm U_0 \leftrightarrow \mp U_0$ ), yielding the relation

$$T_{\sigma, \sigma'}(E, \pm U_0) = T_{\sigma', \sigma}(E, \mp U_0)$$

between the spin-resolved transmission probabilities in the two rocking situations [85]. Inserting this relation into Eq. (30), we observe that the expression for the net charge transmission  $\Delta T_C$  is zero, resulting in vanishing net charge current. Furthermore it can be used to simplify the expression for the net spin current:

$$\begin{aligned} \langle I_S(U_0) \rangle &= \frac{G_S}{e} \int_{E_C}^{\infty} dE \Delta f(E; U_0) \\ &\quad \times [T_{+, -}(E, +U_0) - T_{-, +}(E, +U_0)]. \end{aligned}$$

At  $U_0 = 0$  the Hamiltonian (29) reduces to Eq. (24), which is invariant under the operation of Eq. (26), yielding  $T_{+, -}(E, 0) = T_{-, +}(E, 0)$ . This absence of net spin current  $\langle I_S(U_0 = 0) \rangle = 0$  in the linear response regime is in agreement with the theoretical prediction [85]. However, for finite rocking voltages  $U_0 \neq 0$  the additional potential introduced via  $H_R$  breaks this symmetry and therefore enables different spin flip transmissions and thus a resulting net spin current.

We now turn our attention to the quantum wire shown in Fig. 12a. For this system, which exhibits the symmetries of Eq. (31), we perform numerical calculations at finite bias  $U_0$ , in order to confirm its operability as a spin ratchet. In general, the function  $g(x, y; U_R)$  has to be obtained by self-consistently solving the Schrödinger and Poisson equation of the system. However, in the present treatment we make use of a simple model for  $g(x, y; U_R)$  assuming a linear voltage drop in the region where the point contact is formed in the quantum wire:

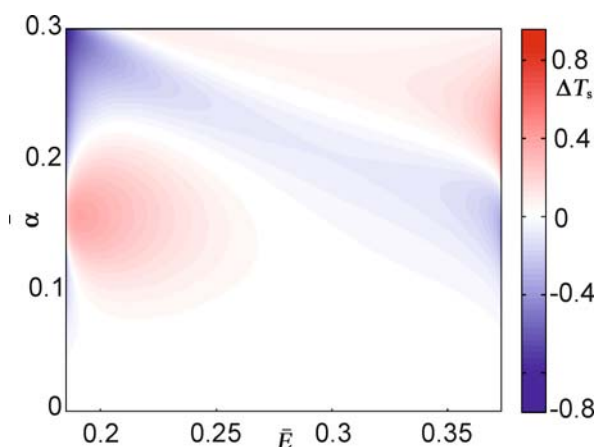
$$g(x, y; U_R) = g(x) = \begin{cases} \frac{1}{2} & \text{for } x < -\frac{L_{PC}}{2}, \\ -\frac{x}{L_{PC}} & \text{for } -\frac{L_{PC}}{2} < x < \frac{L_{PC}}{2}, \\ -\frac{1}{2} & \text{for } x > \frac{L_{PC}}{2}. \end{cases}$$

It is well known that a constriction in a quantum wire acts as an effective potential barrier constituting a region where the voltage applied across the wire is likely to drop. Since the bias voltages we consider are small compared to the energy shift introduced by the constriction, this assumption of a linear voltage drop should be an appropriate approximation for the actual distribution of the electrostatic potential in this wire [104]. Self-consistent calculations have confirmed that this approximation is valid [105].

In Fig. 15 we plot the net spin transmission  $\Delta T_S$  as a function of  $\tilde{\alpha}$  and the injection energy in the range, where both leads support two transversal modes. We observe that  $|\Delta T_S|$  reaches values of up to 0.9 in the parameter range shown. Furthermore, for a given value of injection energy, the sign of  $\Delta T_S$  can be switched by changing  $\alpha_C$ . Since the strength of the Rashba SO interaction  $\alpha_C$  can be tuned via gate voltages [25] the presented system offers also the possibility for experimentally steering and switching the spin current direction.

## Future Directions

In the present work we outlined general theoretical and computational concepts of coherent spin-dependent transport at low temperatures and focussed, with regard to numerical examples and possible experimental realiza-



Spin-Polarized Quantum Transport in Mesoscopic Conductors, Figure 15

Net ratchet spin transmission probability  $\Delta T_S(U_0) = 2(T_{+,-}(+U_0) - T_{-,+}(+U_0))$  for a stripe with a junction (see text) presented as a function of injection energy  $\bar{E}$  and SO interaction strength  $\alpha$  at finite applied voltage  $eU_0 = 0.02\hbar^2/(2m^*a^2)$ . Note the sign change in the spin transmission upon tuning the SO coupling

tions, onto ballistic two-dimensional nanostructures based on non-magnetic high-mobility semiconductors.

In order to experimentally achieve high spin polarizations and reasonable spin currents, if possible at room temperature, broad efforts are made to investigate and design novel materials for spintronics. Here, prominent and promising examples, both with respect to fundamental physics and possible applications, are magnetic semiconductors such as GaMnAs [106] or semimagnetic materials with huge  $g$ -factors, for instance HgTe [107]. Charge transport in these materials is based on holes. However, relatively few theoretical papers deal with phase coherence effects for hole (spin) transport, though the rich band structure and the interplay between heavy and light hole (or electron- and hole-like) degree of freedoms promise interesting additional features.

The theoretical methods for quantum transport, presented here in the context of mesoscopic systems, are also applied and extended to treat transport in a further prospective field, namely through single-molecule junctions [108], for instance (break) junctions with a molecule bridging the gap between two leads or scanning tunneling microscope measurements of tunnel current through molecules at surfaces. In *Molecular Spintronics* [44,109] spin effects in transport through molecules are addressed. This subfield of spin electronics is still in its infancy. On the computational side these systems pose considerable problems since an adequate approach requires an appro-

priate description of the electronic and possibly vibrational properties of the molecule including the coupling to and effects of the leads. Whether (spin) DFT calculations for such an embedded molecule, combined with Landauer-type transport calculations, are appropriate, remains to be an issue, in particular if charging or non-equilibrium effects are involved.

As a further future direction we expect that spin transport in graphene, monolayers of graphite, may evolve as another future research line. After its experimental discovery in 2004 [110], graphene has gained much experimental and huge theoretical attention owing to its many exotic properties such as the massless charge carriers, internal spin-like degree of freedoms and unconventional transport characteristics. Also first experiments on graphene-based nanoconductors, e.g., measurements of the Aharonov–Bohm effect in graphene rings [111], are on their way. Graphene is also viewed as a prospective candidate for spin-electronics, since the spin decoherence and spin relaxation times in graphene are expected to be long [112,113]. Recent promising experiments already succeeded in injecting spins from ferromagnetic metallic contacts into graphene, although the conductance mismatch between graphene and the ferromagnetic leads is expected to suppress the efficiency. Recent theoretical proposals predict efficient spin injection into bulk graphene from graphene ribbons employing the occurrence of current-carrying spin-polarized edge states in the ribbons [114].

## Acknowledgments

Parts of the results presented are based on previous work together with D. Bercioux, D. Frustaglia, and A. Pfund. We further thank I. Adagideli, M. Grifoni and D. Weiss for useful discussions. Two of us, KR and MW, acknowledge financial support by the Deutsche Forschungsgemeinschaft (SFB 689) and MS through the Studienstiftung des Deutschen Volkes.

## Bibliography

### Primary Literature

1. Datta S (2002) *Electronic Transport in Mesoscopic Systems*. Cambridge University Press, Cambridge
2. Rashba EI (1960) Properties of semiconductors with an extremum loop .1. cyclotron and combinational resonance in a magnetic field perpendicular to the plane of the loop. *Sov Phys Solid State* 2:1109–1122
3. Bychkov YA, Rashba EI (1984) Oscillatory effects and the magnetic susceptibility of carriers in inversion layers. *J Phys C: Solid State Phys* 17:6039–6045

4. Hentschel M, Schomerus H, Frustaglia D, Richter K (2004) Aharonov–Bohm physics with spin. I. Geometric phases in one-dimensional ballistic rings. *Phys Rev B* 69:155326
5. Cohen G, Hod O, Rabani E (2007) Constructing spin interference devices from nanometric rings. *Phys Rev B* 76:235120
6. Loss D, Goldbart P, Balatsky AV (1990) Berry's phase and persistent charge and spin currents in textured mesoscopic rings. *Phys Rev Lett* 65:1655–1658
7. Stern A (1992) Berry's phase, motive forces, and mesoscopic conductivity. *Phys Rev Lett* 68:1022–1025
8. Aronov AG, Lyanda-Geller YB (1993) Spin-orbit Berry phase in conducting rings. *Phys Rev Lett* 70:343–346
9. Qian TZ, Su ZB (1994) Spin-orbit interaction and Aharonov–Anandan phase in mesoscopic rings. *Phys Rev Lett* 72: 2311–2315
10. Frustaglia D, Hentschel M, Richter K (2001) Quantum transport in nonuniform magnetic fields: Aharonov–Bohm ring as a spin switch. *Phys Rev Lett* 87:256602
11. Frustaglia D, Hentschel M, Richter K (2004) Aharonov–Bohm physics with spin. II. Spin-flip effects in two-dimensional ballistic systems. *Phys Rev B* 69:155327
12. Nitta J, Bergsten T (2007) Time reversal Aharonov–Casher effect using Rashba spin-orbit interaction. *New J Phys* 9:341
13. Murgio AF, Heida JP, Klapwijk TM, van Wees BJ, Borghs G (1998) Ensemble-average spectrum of Aharonov–Bohm conductance oscillations: Evidence for spin-orbit-induced Berry's phase. *Phys Rev Lett* 80:1050–1053
14. Yau JB, De Poortere EP, Shayegan M (2002) Aharonov–Bohm oscillations with spin: Evidence for Berry's phase. *Phys Rev Lett* 88:146801
15. König M, Tschetschekin A, Hankiewicz EM, Sinova J, Hock V, Daumer V, Schäfer M, Becker CR, Buhmann H, Molenkamp LW (2006) Direct observation of the Aharonov–Casher phase. *Phys Rev Lett* 96:076804
16. Grbić B, Leturcq R, Ihn T, Ensslin K, Reuter D, Wieck AD (2007) Aharonov–Bohm oscillations in the presence of strong spin-orbit interactions. *Phys Rev Lett* 99:176803
17. Frustaglia D, Richter K (2004) Spin interference effects in ring conductors subject to Rashba coupling. *Phys Rev B* 69:235310
18. Nitta J, Meijer FE, Takayanagi H (1999) Spin-interference device. *Appl Phys Lett* 75:695–697
19. Mal'shukov AG, Shlyapin VV, Chao KA (1999) Effect of the spin-orbit geometric phase on the spectrum of Aharonov–Bohm oscillations in a semiconductor mesoscopic ring. *Phys Rev B* 60:R2161–R2164
20. Splettstoesser J, Governale M, Zülicke U (2003) Persistent current in ballistic mesoscopic rings with Rashba spin-orbit coupling. *Phys Rev B* 68:165341
21. Souma S, Nikolić BK (2005) Spin Hall current driven by quantum interferences in mesoscopic Rashba rings. *Phys Rev Lett* 94:106602
22. Popp M, Frustaglia D, Richter K (2003) Spin filter effects in mesoscopic ring structures. *Nanotechnology* 14:347–351
23. Bellucci S, Onorato P (2007) Filtering of spin currents based on a ballistic ring. *J Phys: Condens Matter* 19:395020
24. Ionicioiu R, D'Amico I (2003) Mesoscopic Stern–Gerlach device to polarize spin currents. *Phys Rev B* 67:041307
25. Nitta J, Akazaki T, Takayanagi H, Enoki T (1997) Gate control of spin-orbit interaction in an inverted  $\text{In}_{0.53}\text{Ga}_{0.47}\text{As}/\text{In}_{0.52}\text{Al}_{0.48}\text{As}$  heterostructure. *Phys Rev Lett* 78:1335–1338
26. Schmidt G, Ferrand D, Molenkamp LW, Filip AT, van Wees BJ (2000) Fundamental obstacle for electrical spin injection from a ferromagnetic metal into a diffusive semiconductor. *Phys Rev B* 62:R4790–R4793
27. Sinova J, Culcer D, Niu Q, Sinitsyn NA, Jungwirth T, MacDonald AH (2004) Universal intrinsic spin Hall effects. *Phys Rev Lett* 92:126603
28. New J Phys (2007) Focus on Spintronics in Reduced Dimensions. *New J Phys* 9
29. Corresponding settings, where a precessing magnetization in a ferromagnet emits spin currents, were reviewed by Tserkovnyak Y, Brataas A, Bauer GEW, Halperin BI (2005) Non-local magnetization dynamics in ferromagnetic heterostructures. *Rev Mod Phys* 77:1375
30. Landauer R (1957) Spatial variation of currents and fields due to localized scatterers in metallic conduction. *IBM J Res Dev* 1:223–231
31. Büttiker M, Imry Y, Landauer R, Pinhas S (1985) Generalized many-channel conductance formula with application to small rings. *Phys Rev B* 31:6207–6215
32. Stone AD, Szafer A (1988) What is measured when you measure a resistance? – The Landauer formula revisited. *IBM J Res Dev* 32:384–413
33. Ferry DK, Goodnick SM (2001) Transport in Nanostructures. Cambridge University Press, Cambridge
34. Fisher DS, Lee PA (1981) Relation between conductivity and transmission matrix. *Phys Rev B* 23:6851–6854
35. Baranger HU, Stone AD (1989) Electrical linear-response theory in an arbitrary magnetic field: A new fermi-surface formation. *Phys Rev B* 40:8169–8193
36. Kimball GE, Shortley GH (1934) The numerical solution of Schrödinger's equation. *Phys Rev* 45:815–820
37. Pauling L, Wilson EB (1935) Introduction to Quantum Mechanics. Dover, New York
38. Havu P, Havu V, Puska MJ, Nieminen RM (2004) Nonequilibrium electron transport in two-dimensional nanostructures modeled using Green's functions and the finite-element method. *Phys Rev B* 69:115325
39. Bowen RC, Klimeck G, Lake RK, Frensley WR, Moise T (1997) Quantitative simulation of a resonant tunneling diode. *J Appl Phys* 81:3207–3213
40. Sanvito S, Lambert CJ, Jefferson JH, Bratkovsky AM (1999) General Green's-function formalism for transport calculations with *spd* Hamiltonians and giant magnetoresistance in Co- and Ni-based magnetic multilayers. *Phys Rev B* 59:11936–11948
41. Luisier M, Schenk A, Fichtner W, Klimeck G (2006) Atomistic simulation of nanowires in the  $sp^3d^5s^*$  tight-binding formalism: From boundary conditions to strain calculations. *Phys Rev B* 74:205323
42. Brandbyge M, Mozos JL, Ordejón P, Taylor J, Stokbro K (2002) Density-functional method for nonequilibrium electron transport. *Phys Rev B* 65:165401
43. Di Carlo A, Pecchia A, Latessa L, Frauenheim T, Seifert G (2005) Tight-binding DFT for molecular electronics (gDFTB). In: Cuniberti G, Fagas G, Richter K (eds) *Introducing Molecular Electronics*. Springer, Berlin, pp 153–184
44. Rocha AR, García-Suárez VM, Bailey S, Lambert C, Ferrer J, Sanvito S (2006) Spin and molecular electronics in atomically generated orbital landscapes. *Phys Rev B* 73:085414



45. Lopez Sancho MP, Lopez Sancho JM, Rubio J (1984) Quick iterative scheme for the calculation of transfer matrices: application to Mo (100). *J Phys F: Met Phys* 14:1205–1215
46. Lopez Sancho MP, Lopez Sancho JM, Rubio J (1985) Highly convergent schemes for the calculation of bulk and surface Green functions. *J Phys F: Met Phys* 15:851–858
47. Krstić PS, Zhang XG, Butler WH (2002) Generalized conductance formula for the multiband tight-binding model. *Phys Rev B* 66:205319
48. Usuki T, Takatsu M, Kiehl RA, Yokoyama N (1994) Numerical analysis of electron-wave detection by a wedge-shaped point contact. *Phys Rev B* 50:7615–7625
49. Usuki T, Saito M, Takatsu M, Kiehl RA, Yokoyama N (1995) Numerical analysis of ballistic-electron transport in magnetic fields by using a quantum point contact and a quantum wire. *Phys Rev B* 52:8244–8255
50. Lambert CJ, Weaire D (1980) Decimation and Anderson localization. *Phys Status Solid B* 101:591–595
51. Leadbeater M, Lambert CJ (1998) A decimation method for studying transport properties of disordered systems. *Ann Phys* 7:498–502
52. Mamaluy D, Vasileska D, Sabathil M, Zibold T, Vogl P (2005) Contact block reduction method for ballistic transport and carrier densities of open nanostructures. *Phys Rev B* 71:245321
53. Thouless DJ, Kirkpatrick S (1981) Conductivity of the disordered linear chain. *J Phys C: Solid State Phys* 14:235–245
54. Lee PA, Fisher DS (1981) Anderson localization in two dimensions. *Phys Rev Lett* 47:882–885
55. MacKinnon A (1985) The calculation of transport properties and density of states of disordered solids. *Z Phys B* 59:385–390
56. Baranger HU, DiVincenzo DP, Jalabert RA, Stone AD (1991) Classical and quantum ballistic-transport anomalies in microjunctions. *Phys Rev B* 44:10637–10675
57. Lake R, Klimeck G, Bowen RC, Jovanovic D (1997) Single and multiband modeling of quantum electron transport through layered semiconductor devices. *J Appl Phys* 81:7845–7869
58. Lassl A, Schlagheck P, Richter K (2007) Effects of short-range interactions on transport through quantum point contacts: A numerical approach. *Phys Rev B* 75:045346
59. Drouvelis P, Schmelcher P, Bastian P (2006) Parallel implementation of the recursive Green's function method. *J Comp Phys* 215:741–756
60. Rotter S, Tang JZ, Wirtz L, Trost J, Burgdörfer J (2000) Modular recursive Green's function method for ballistic quantum transport. *Phys Rev B* 62:1950–1960
61. Rotter S, Weingartner B, Rohringer N, Burgdörfer J (2003) Ballistic quantum transport at high energies and high magnetic fields. *Phys Rev B* 68:165302
62. Cuthill E, McKee J (1969) Reducing the bandwidth of sparse symmetric matrices. In: *Proc 24th Nat Conf, ACM, New York*, pp 157–172
63. George A (1971) Computer implementation of the finite element method. *Tech Rep STAN-CS-71-208, Computer Sci Dept, Stanford Univ, Stanford*
64. Liu WH, Sherman AH (1976) Comparative analysis of the Cuthill–McKee and the reverse Cuthill–McKee ordering algorithms for sparse matrices. *SIAM J Num Anal* 13:198–213
65. Gibbs NE, William G, Poole J, Stockmeyer PK (1976) An algorithm for reducing the bandwidth and profile of a sparse matrix. *SIAM J Num Anal* 13:236–250
66. Wimmer M, Richter K (2008) Optimal block-tridiagonalization of matrices for coherent charge transport. *arXiv:0806.2739v1*
67. Whaley RC, Petitet A, Dongarra JJ (2001) Automated empirical optimization of software and the ATLAS project. *Parallel Comput* 27:3–35
68. van Houten H, Beenakker CWJ, Williamson JG, Broekaart MEI, van Loosdrecht PHM, van Wees BJ, Mooij JE, Foxon CT, Harris JJ (1989) Coherent electron focusing with quantum point contacts in a two-dimensional electron gas. *Phys Rev B* 39:8556–8575
69. Usaj G, Balseiro CA (2004) Transverse electron focusing in systems with spin-orbit coupling. *Phys Rev B* 70:041301
70. Govorov AO, Kalameitsev AV, Dulka JP (2004) Spin-dependent transport of electrons in the presence of a smooth lateral potential and spin-orbit interaction. *Phys Rev B* 70:245310
71. Reynoso A, Usaj G, Balseiro CA (2007) Detection of spin polarized currents in quantum point contacts via transverse electron focusing. *Phys Rev B* 75:085321
72. Rokhinson LP, Larkina V, Lyanda-Geller YB, Pfeiffer LN, West KW (2004) Spin separation in cyclotron motion. *Phys Rev Lett* 93:146601
73. Potok RM, Folk JA, Marcus CM, Umansky V (2002) Detecting spin-polarized currents in ballistic nanostructures. *Phys Rev Lett* 89:266602
74. Folk JA, Potok RM, Marcus CM, Umansky V (2003) A Gate-Controlled Bidirectional Spin Filter Using Quantum Coherence. *Science* 299:679–682
75. Rokhinson LP, Pfeiffer LN, West KW (2006) Spontaneous spin polarization in quantum point contacts. *Phys Rev Lett* 96:156602
76. Fabian J, Das Sarma S (2002) Spin transport in inhomogeneous magnetic fields: A proposal for Stern–Gerlach-like experiments with conduction electrons. *Phys Rev B* 66:024436
77. Kiselev AA, Kim KW (2001) T-shaped ballistic spin filter. *Appl Phys Lett* 78:775–777
78. Ohe JI, Yamamoto M, Ohtsuki T, Nitta J (2005) Mesoscopic Stern–Gerlach spin filter by nonuniform spin-orbit interaction. *Phys Rev B* 72:041308
79. Cummings AW, Akis R, Ferry DK (2006) Electron spin filter based on Rashba spin-orbit coupling. *Appl Phys Lett* 89:172115
80. Eto M, Hayashi T, Kurotani Y (2005) Spin polarization at semiconductor point contacts in absence of magnetic field. *J Phys Soc Jpn* 74:1934
81. Zhai F, Xu HQ (2007) Spin filtering and spin accumulation in an electron stub waveguide with spin-orbit interaction. *Phys Rev B* 76:035306
82. Scheid M, Pfund A, Bercieux D, Richter K (2007) Coherent spin ratchets: A spin-orbit based quantum ratchet mechanism for spin-polarized currents in ballistic conductors. *Phys Rev B* 76:195303
83. Song JF, Ochiai Y, Bird JP (2003) Fano resonances in open quantum dots and their application as spin filters. *Appl Phys Lett* 82:4561–4563
84. Zhai F, Xu HQ (2006) Spin filtering in single magnetic barrier structures revisited. *Appl Phys Lett* 88:032502



85. Scheid M, Bercieux D, Richter K (2007) Zeeman ratchets: pure spin current generation in mesoscopic conductors with non-uniform magnetic fields. *New J Phys* 9:401
86. Shi QW, Zhou J, Wu MW (2004) Spin filtering through a double-bend structure. *Appl Phys Lett* 85:2547–2549
87. Zhai F, Xu HQ (2005) Generation of spin polarization in two-terminal electron waveguides by spin-orbit interaction and magnetic field modulations. *Phys Rev B* 72:085314
88. Shi J, Zhang P, Xiao D, Niu Q (2006) Proper definition of spin current in spin-orbit coupled systems. *Phys Rev Lett* 96:076604
89. Zhai F, Xu HQ (2005) Symmetry of spin transport in two-terminal waveguides with a spin-orbital interaction and magnetic field modulations. *Phys Rev Lett* 94:246601
90. Silvestrov PG, Mishchenko EG (2006) Polarized electric current in semiclassical transport with spin-orbit interaction. *Phys Rev B* 74:165301
91. van Wees BJ, van Houten H, Beenakker CWJ, Williamson JG, Kouwenhoven LP, van der Marel D, Foxon CT (1988) Quantized conductance of point contacts in a two-dimensional electron gas. *Phys Rev Lett* 60:848–850
92. Wharam DA, Thornton TJ, Newbury R, Pepper M, Ahmed H, Frost JEF, Hasko DG, Peacock DC, Ritchie DA, Jones GAC (1988) One-dimensional transport and the quantisation of the ballistic resistance. *J Phys C: Solid State Phys* 21:L209–L214
93. Sánchez D, Serra L (2006) Fano–Rashba effect in a quantum wire. *Phys Rev B* 74:153313
94. Dresselhaus G (1955) Spin-orbit coupling effects in zinc blende structures. *Phys Rev* 100:580–586
95. Bardarson JH, Adagideli I, Jacquod P (2007) Mesoscopic spin Hall effect. *Phys Rev Lett* 98:196601
96. Hankiewicz EM, Molenkamp LW, Jungwirth T, Sinova J (2004) Manifestation of the spin Hall effect through charge-transport in the mesoscopic regime. *Phys Rev B* 70:241301
97. Sheng L, Sheng DN, Ting CS (2005) Spin-Hall effect in two-dimensional electron systems with Rashba spin-orbit coupling and disorder. *Phys Rev Lett* 94:016602
98. Governale M, Taddei F, Fazio R (2003) Pumping spin with electrical fields. *Phys Rev B* 68:155324
99. Sharma P, Brouwer PW (2003) Mesoscopic effects in adiabatic spin pumping. *Phys Rev Lett* 91:166801
100. Mucciolo ER, Chamon C, Marcus CM (2002) Adiabatic quantum pump of spin-polarized current. *Phys Rev Lett* 89:146802
101. Watson SK, Potok RM, Marcus CM, Umansky V (2003) Experimental realization of a quantum spin pump. *Phys Rev Lett* 91:258301
102. Reimann P (2002) Brownian motors: noisy transport far from equilibrium. *Phys Rep* 361:57–265
103. Scheid M, Wimmer M, Bercieux D, Richter K (2006) Zeeman ratchets for ballistic spin currents. *Phys Status Solidi (c)* 3:4235
104. McLennan MJ, Lee Y, Datta S (1991) Voltage drop in mesoscopic systems: A numerical study using a quantum kinetic equation. *Phys Rev B* 43:13846–13884
105. Lassl A, Scheid M, Richter K (2008) Unpublished
106. Ohno H, Shen A, Matsukura F, Oiwa A, Endo A, Katsumoto S, Iye Y (1996) (Ga,Mn)As: A new diluted magnetic semiconductor based on GaAs. *Appl Phys Lett* 69:363–365
107. König M, Wiedmann S, Brüne C, Roth A, Buhmann H, Molenkamp LW, Qi XL, Zhang SC (2007) Quantum spin Hall insulator state in HgTe quantum wells. *Science* 318:766–770
108. Cuniberti G, Fagas G, Richter K (eds) (2005) *Introducing Molecular Electronics*. Springer, Berlin
109. Emberly E, Kirczenow G (2002) Molecular spintronics: spin-dependent electron transport in molecular wires. *Chemical Physics* 281:311–324
110. Novoselov KS, Geim AK, Morozov SV, Jiang D, Zhang Y, Dubonos SV, Grigorieva IV, Firsov AA (2004) Electric field effect in atomically thin carbon films. *Science* 306:666–669
111. Russo S, Oostinga JB, Wehenkel D, Heersche HB, Sobhani SS, Vandersypen LMK, Morpurgo AF (2007) Aharonov–Bohm effect in graphene. *arXiv:0711.1508v1*
112. Huertas-Hernando D, Guinea F, Brataas A (2006) Spin-orbit coupling in curved graphene, fullerenes, nanotubes, and nanotube caps. *Phys Rev B* 74:155426
113. Min H, Hill JE, Sinitsyn NA, Sahu BR, Kleinman L, MacDonald AH (2006) Intrinsic and Rashba spin-orbit interactions in graphene sheets. *Phys Rev B* 74:165310
114. Wimmer M, Adagideli I, Berber S, Tománek D, Richter K (2008) Spin transport in rough graphene nanoribbons. *Phys Rev Lett* 100:177207

### Books and Reviews

- Bruus H, Flensberg K (2004) *Many-body Quantum Theory in Condensed Matter Physics: An Introduction*. Oxford University Press, Oxford
- Datta S (2002) *Electronic Transport in Mesoscopic Systems*. Cambridge University Press, Cambridge
- Fabian J, Matos-Abiadue A, Ertler C, Stano P, Žutić I (2007) *Semiconductor Spintronics*. *Acta Physica Slovaca* 57:565–907
- Ferry DK, Goodnick SM (2001) *Transport in Nanostructures*. Cambridge University Press, Cambridge

## Stability and Feedback Stabilization

EDUARDO D. SONTAG

Department of Mathematics, Rutgers University,  
New Brunswick, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Linear Systems](#)

[Nonlinear Systems: Continuous Feedback](#)

[Discontinuous Feedback](#)

[Sensitivity to Small Measurement Errors](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Stability** A globally asymptotically stable equilibrium is a state with the property that all solutions converge to this state, with no large excursions.

**Stabilization** A system is stabilizable (with respect to a given state) if it is possible to find a feedback law that renders that state a globally asymptotically stable equilibrium.

**Lyapunov and control-Lyapunov functions** A control-Lyapunov function is a scalar function which decreases along trajectories, if appropriate control actions are taken. For systems with no controls, one has a Lyapunov function.

## Definition of the Subject

The problem of stabilization of equilibria is one of the central issues in control. In addition to its intrinsic interest, it represents a first step towards the solution of more complicated problems, such as the stabilization of periodic orbits or general invariant sets, or the attainment of other control objectives, such as tracking, disturbance rejection, or output feedback, all of which may be interpreted as requiring the stabilization of some quantity (typically, some sort of “error” signal). A very special case, when there are no inputs, is that of stability.

## Introduction

This article discusses the problem of stabilization of an equilibrium, which we take without loss of generality to be the origin, for a finite-dimensional system  $\dot{x} = f(x, u)$ . The objective is to find a *feedback law*  $u = k(x)$  which renders the origin of the “closed-loop” system  $\dot{x} = f(x, k(x))$  globally asymptotically stable. The problem of stabilization of equilibria is one of the central issues in control. In addition to its intrinsic interest, it represents a first step towards the solution of more complicated problems, such as the stabilization of periodic orbits or general invariant sets, or the attainment of other control objectives, such as tracking, disturbance rejection, or output feedback, all of which may be interpreted as requiring the stabilization of some quantity (typically, some sort of “error” signal). A very special case (when there are no inputs  $u$ ) is that of stability.

After setting up the basic definitions, we consider *linear* systems. Linear systems are widely used as models for physical processes, and they also play a major role in the general theory of local stabilization. We briefly review pole assignment and linear-quadratic optimization as approaches to obtaining feedback stabilizers.

In general, there is a close connection between the existence of continuous stabilizing feedbacks and smooth *control-Lyapunov functions*, (clf’s), which constitute an extension of the classical concept of Lyapunov functions from dynamical system theory. We discuss the role of clfs

in design methods and “universal” formulas for feedback controls.

For nonlinear systems, it has been known since the late 1970s that, in general, there are topological obstructions to the existence of even continuous stabilizers. We review these obstructions, using tools from degree theory.

Finally, we turn to discontinuous stabilization and the associated issue of defining precisely a “solution” for a differential equation with discontinuous right-hand side. We introduce techniques from nonsmooth analysis and differential games, in order to deal with discontinuous controllers. In particular, we discuss the effect of measurement errors on the performance of such controllers.

## Preliminaries

In this article, we restrict attention to continuous-time deterministic systems whose states evolve in finite-dimensional Euclidean spaces  $\mathbb{R}^n$ . (This excludes many other equally important objects of study in control theory: systems which evolve on infinite dimensional spaces and are described by PDE’s, systems evolving on manifolds which serve to model state constraints, discrete-time systems described by difference equations, and stochastic systems, among others.) In order to streamline the presentation, we suppose throughout that controls take values in  $\mathcal{U} = \mathbb{R}^m$  (constraints in controls would lead to proper subsets  $\mathcal{U}$ ). A *control* (other names: *input*, *forcing function*) is any measurable locally essentially bounded map  $u(\cdot): [0, \infty) \rightarrow \mathcal{U} = \mathbb{R}^m$ . In general, we use the notation  $|x|$  for Euclidean norms, and use  $\|u\|$  to indicate the essential supremum of a function  $u(\cdot)$ . For basic terminology and facts about control systems, see [25].

Given a map  $f: \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$  which is locally Lipschitz and satisfies  $f(0, 0) = 0$ , we consider the associated forced system of ordinary differential equations

$$\dot{x}(t) = f(x(t), u(t)). \quad (1)$$

The maximal solution  $x(\cdot)$  of (1) which corresponds to a given initial state  $x(0) = x^0$  and to a given control  $u$  is defined on some maximal interval  $[0, t_{\max}(x^0, u))$ , and is denoted by  $x(t, x^0, u)$ . In the special case when  $f$  does not depend on  $u$ , we have an unforced system, or system with no inputs

$$\dot{x}(t) = f(x(t)). \quad (2)$$

Unforced systems are associated to a controlled system (1) in two different ways. The first is when one substitutes a feedback law  $u = k(x)$  in (1) to obtain a “closed-loop” system  $\dot{x} = f(x, k(x))$ . The second is when one considers

instead the autonomous system  $\dot{x} = f(x, 0)$  which models the behavior of (1) in the absence of any controls. All definitions stated for unforced systems are implicitly applied also to systems with inputs (1) by setting  $u \equiv 0$ ; for instance, we define the global asymptotic stability (GAS) property for (2), but we say that (1) is GAS if  $\dot{x} = f(x, 0)$  is. For systems with no inputs (2) we write just  $x(t, x^0)$  instead of  $x(t, x^0, u)$ .

**Stability and Asymptotic Controllability** Stability is one of the most important objectives in control theory, because a great variety of problems can be recast in stability terms. This includes questions of driving a system to a desired configuration (e.g., an inverted pendulum on a cart, to its upwards position), or the problem of tracking a reference signal (such as a pilot's command to an aircraft). We focus in this talk on global asymptotic stabilization.

Recall that the class of  $\mathcal{KL}$  functions consists of all  $\alpha: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$  which are continuous, strictly increasing, unbounded, and satisfy  $\alpha(0) = 0$ . The class of  $\mathcal{KL}$  functions consists of those  $\beta: \mathbb{R}_{\geq 0} \times \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$  with the properties that

- (1)  $\beta(\cdot, t) \in \mathcal{KL}$  for all  $t$ , and
- (2)  $\beta(r, t)$  decreases to zero as  $t \rightarrow \infty$ .

We will also use  $\mathcal{N}$  to denote the set of all nondecreasing functions  $\sigma: \mathbb{R}_{\geq 0} \rightarrow \mathbb{R}_{\geq 0}$ . Expressed in terms of such comparison functions, the property of *global asymptotic stability* (GAS) of the origin for a system with no inputs (2) becomes:

$$(\exists \beta \in \mathcal{KL}) \quad |x(t, x^0)| \leq \beta(|x^0|, t) \quad \forall x^0, \forall t \geq 0.$$

This definition is equivalent to a more classical “ $\varepsilon$ - $\delta$ ” definition usually provided in textbooks, which defines GAS as the combination of stability and global attractivity. For one implication, simply observe that

$$|x(t, x^0)| \leq \beta(|x^0|, 0)$$

provides the stability (or “small overshoot”) property, while

$$|x(t, x^0)| \leq \beta(|x^0|, t) \xrightarrow{t \rightarrow \infty} 0$$

gives attractivity. The converse implication is an easy exercise.

More generally, we define what it means for the system with inputs (1) to be (open loop, globally) *asymptotically controllable* (AC) (to the origin). The definition amounts to requiring that for each initial state  $x^0$  there exists some control  $u = u_{x^0}(\cdot)$  defined on  $[0, \infty)$ , such

that the corresponding solution  $x(t, x^0, u)$  is defined for all  $t \geq 0$ , and converges to zero as  $t \rightarrow \infty$ , with “small” overshoot. Moreover, we wish to rule out the possibility that  $u(t)$  becomes unbounded for  $x$  near zero. The precise formulation is as follows.

$$(\exists \beta \in \mathcal{KL})(\exists \sigma \in \mathcal{N}) \quad \forall x^0 \in \mathbb{R}^n \exists u(\cdot), \|u\| \leq \sigma(|x^0|),$$

$$|x(t, x^0, u)| \leq \beta(|x^0|, t) \quad \forall t \geq 0.$$

In particular, (global) asymptotic stability amounts to the existence of  $\beta \in \mathcal{KL}$  such that  $|x(t, x^0, u)| \leq \beta(|x^0|, t)$  holds for all  $t \geq 0$ . A very special case is that of *exponential* stability, in which an estimate of the type  $|x(t, x^0, u)| \leq Me^{-\lambda t}|x^0|$  holds. For linear systems (see below), asymptotic stability and exponential stability coincide. It is a puzzling fact that for general systems, one can find continuous coordinate changes that make asymptotically stable systems exponentially stable [8, 13] (a fact of little practical utility, since finding such coordinate changes is as hard as establishing stability to being with).

**Feedback Stabilization** A map  $k: \mathbb{R}^n \rightarrow \mathcal{U}$  is a *feedback stabilizer* for the system with inputs (1) if  $k$  is locally bounded (that is,  $k$  is bounded on each bounded subset of  $\mathbb{R}^n$ ),  $k(0) = 0$ , and the closed-loop system

$$\dot{x} = f(x, k(x)) \tag{3}$$

is GAS, i.e. there is some  $\beta \in \mathcal{KL}$  so that  $|x(t)| \leq \beta(|x(0)|, t)$  for all solutions and all  $t \geq 0$ . (A technical difficulty with this definition lies the possible lack of solutions of (1) when  $k$  is not regular enough. We ignore this for now, but will most definitely return to this issue later.)

For example, if (1) is a model of an undamped spring/mass system, where  $u$  represents the net effect of external forces, one obvious way to asymptotically stabilize the system is to introduce damping. In control-theoretic terms, this means that we choose  $u(t) = k(x(t))$  to be a negative linear function of the velocity. Physically, one may implement a feedback controller by means of an analog device. In the example of the spring/mass system, one could achieve this by adding friction or connecting a dashpot. Alternatively, in modern control technology, one uses a digital computer to measure the state  $x$  and compute the appropriate control action to be applied. (There are many implementation issues which arise in digital control and are ignored in our theoretical formulation “ $u = k(x)$ ”, among them the effect of delays in the actual computation of the control  $u(t)$  to be applied at time  $t$ , and the effect of quantization due to the finite precision of measuring devices and the digital nature of the computer. These issues

are addressed in the literature, although a comprehensive theoretical framework is still lacking.)

Observe that, obviously, if there exists a feedback stabilizer for (1), then (1) is also AC (we just use  $u(t) := k(x(t, x^0))$  as  $u_{x^0}$ ). Thus, it is very natural to ask whether the converse holds: *is every asymptotically controllable system also feedback stabilizable?*

## Linear Systems

A *linear system* is a system (1) for which the map  $f$  is linear. In other words, there are two linear transformations  $A: \mathbb{R}^n \rightarrow \mathbb{R}^n$  and  $B: \mathbb{R}^m \rightarrow \mathbb{R}^n$  so that the equations take the form

$$\dot{x} = Ax + Bu. \quad (4)$$

Such a system is completely specified once that we are given  $A$  and  $B$ , which we identify by abuse of notation with their respective matrices  $A \in \mathbb{R}^{n \times n}$  and  $B \in \mathbb{R}^{n \times m}$  with respect to the canonical bases in  $\mathbb{R}^n$  and  $\mathbb{R}^m$ . We also say “the system  $(A, B)$ ” when referring to (4).

It is natural to look specifically for *linear* feedbacks  $k: \mathbb{R}^n \rightarrow \mathbb{R}^m$  which stabilize a linear system (just as a linear term, inversely proportional to velocity, stabilizes an undamped harmonic oscillator). (In fact, this is no loss of generality, since it can be easily proved for linear systems [25] that if a feedback stabilizer  $u = k(x)$  exists, then there also exists a linear feedback stabilizer.) We write  $u = Fx$ , when expressing  $k(x) = Fx$  in matrix terms with respect to the canonical bases. Substituting this control law into (4) results in the equation  $\dot{x} = (A + BF)x$ . Thus, the mathematical problem reduces to:

given  $A \in \mathbb{R}^{n \times n}$  and  $B \in \mathbb{R}^{n \times m}$ , find  $F \in \mathbb{R}^{m \times n}$   
such that  $A + BF$  is Hurwitz.

(Recall that a Hurwitz matrix is one all whose eigenvalues have negative real parts, and that the origin of the system  $\dot{x} = Hx$  is globally asymptotically stable if and only if  $H$  is a Hurwitz matrix.) The fundamental stabilization result for linear systems is as follows [25]:

**Theorem 1** *A linear system is asymptotically controllable if and only if it admits a linear feedback stabilizer.*

## A Remark on Linearization

If the dynamics map  $f$  in (1) is continuously differentiable, we may expand to first order  $f(x, u) = Ax + Bu + o(x, u)$ . Let us suppose that the linearized system  $(A, B)$  is AC, and pick a linear feedback stabilizer  $u = Fx$ , whose existence

is guaranteed by Theorem 1. Then, the same feedback law  $k(x) = Fx$ , when fed back into the original system (1), results in  $\dot{x} = f(x, Fx) = (A + BF)x + o(x)$ . Thus,  $k$  locally stabilizes the origin for the nonlinear system. Of course, the assumption that the linearization is AC is not always satisfied. Systems in which inputs enter multiplicatively, such as those controlling reaction rates in chemical problems, lead to degenerate linearizations. In addition, even if the linearized system  $(A, B)$  is AC, in general a linear stabilizer  $u = Fx$  will not work as a global stabilizer. For example, the system  $\dot{x} = x + x^2 + u$  can be locally stabilized with  $u := -2x$ , but any linear feedback  $u = -fx$  ( $f > 1$ ) results in an additional equilibrium away from the origin (at  $x = f - 1$ ). Nonlinear feedback must be used (obviously, in this, example,  $u = -2x - x^2$  works for global stabilization).

Returning to linear systems, let us note that Theorem 1 is of great interest because (1) there is a simple algebraic test to check the AC property, and (2) there are several practically useful algorithms for obtaining a stabilizing  $F$ , including pole placement and optimization, which we sketch next.

## Pole Placement

The first technique for stabilization is purely algebraic. In order to simplify this exposition, we will suppose that the system (4) is not just AC but is in fact *controllable*, meaning that every state can be steered, in finite time, to every other state. (Any AC system (4) can be decomposed into two components, of which one is already GAS and the other one is controllable, cf. [25], so this represents no loss of generality.) Controllability is characterized by the property – generically satisfied for pairs  $(A, B)$  – that

$$\text{rank} \begin{bmatrix} B & AB & A^2B & \dots & A^{n-1}B \end{bmatrix} = n$$

(note that the composite matrix shown has  $n$  rows and  $nm$  columns).

Two pairs  $(A, B)$  and  $(\tilde{A}, \tilde{B})$  are said to be *feedback equivalent* if there exist  $T \in GL(n, \mathbb{R})$ ,  $F_0 \in \mathbb{R}^{m \times n}$ , and  $V \in GL(m, \mathbb{R})$  so that

$$(A + BF_0)T = T\tilde{A} \quad \text{and} \quad BV = T\tilde{B}. \quad (5)$$

Feedback equivalence corresponds to changes of basis in the state and control-value spaces (invertible matrices  $T$  and  $V$ , respectively) and feedback transformations  $u = F_0x + u'$ , where  $u'$  is a new control. (An equivalent way to describe feedback equivalence is by the requirement that two pairs should be in the same orbit un-

der the action of a “feedback group” which is obtained as a suitable semidirect product of  $GL(n, \mathbb{R})$ ,  $(\mathbb{R}^{m \times n}, +)$ , and  $GL(m, \mathbb{R})$ .) Controllability is preserved under feedback equivalence. Moreover, if (5) holds and if one finds a matrix  $\tilde{F}$  so that  $\tilde{A} + \tilde{B}\tilde{F}$  is Hurwitz, then

$$A + BF = T(\tilde{A} + \tilde{B}\tilde{F})T^{-1}$$

is also Hurwitz, where  $F := F_0 + \tilde{V}\tilde{F}T^{-1}$ . Thus, the task of finding a stabilizing feedback  $F$  can be reduced to the same problem for any pair  $(\tilde{A}, \tilde{B})$  which is feedback equivalent to the given pair  $(A, B)$ .

One then proceeds to show that there always exists an equivalent pair  $(\tilde{A}, \tilde{B})$  which has a form simple enough that the existence of  $\tilde{F}$  is trivial to establish. In order to find such a pair, it is useful to study the classification of controllable pairs under feedback equivalence. This classification is closely related to Kronecker’s theory of “matrix pencils” applied to polynomial matrices  $[\lambda I - A, B] = \lambda[I, 0] + [-A, B]$  modulo matrix equivalence, cf. [25]. The orbits under feedback equivalence are in one-to-one correspondence with the possible partitions of  $n = \kappa_1 + \dots + \kappa_r$  into the sum of  $r$  positive integers,  $r \leq m$ , and in each orbit one can find a pair  $(\tilde{A}, \tilde{B})$  which is in “controller canonical form”, for which  $F$  can be trivially found. For simplicity, let us just discuss here the very special case of single-input systems ( $m = 1$ ). For this case, the action of the feedback group is transitive, and each controllable system is feedback equivalent to the following special system:

$$A := \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ 0 & 0 & 0 & \dots & 0 \end{pmatrix} \quad B := \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{pmatrix}.$$

For this system, a stabilizing feedback is trivial to obtain. Indeed, take the polynomial  $p(\lambda) = (\lambda + 1)^n = \lambda^n - \alpha_n \lambda^{n-1} - \dots - \alpha_2 \lambda - \alpha_1$ . With  $F = (\alpha_1, \alpha_2, \alpha_3, \dots, \alpha_n)$ ,

$$A + BF := \begin{pmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \\ \alpha_1 & \alpha_2 & \alpha_3 & \dots & \alpha_n \end{pmatrix}$$

has characteristic polynomial  $(\lambda + 1)^n$ , and hence is a Hurwitz matrix, as required for stabilization.

Observe that, instead of the particular  $p(\lambda)$  which we used, we could have picked any polynomial all whose roots

have negative real parts, and the same argument applies. The conclusion is that, not only can we make  $A + BF$  Hurwitz, but we can assign to it any desired set of  $n$  eigenvalues (as long as they form a set closed under conjugation). This is the reason that the technique is called *eigenvalue placement* (or “pole placement” because the eigenvalues of  $A$  are the poles of the resolvent  $(\lambda I - A)^{-1}$ ). See Chap. 5 in [25] for a detailed treatment of the pole placement problem.

### Variational Approach

A second technique for stabilization is based on optimal control techniques. We first pick any two symmetric positive definite matrices  $R \in \mathbb{R}^{m \times m}$  and  $Q \in \mathbb{R}^{n \times n}$  (for instance the identity matrices of the respective sizes). Next, we consider the problem of minimizing, for each initial state  $x^0$  at time  $t = 0$ , the infinite-horizon cost

$$J_{x^0}(u) := \int_0^\infty \{u(t)'Ru(t) + x(t)'Qx(t)\} dt$$

over all controls  $u: [0, \infty) \rightarrow \mathbb{R}^m$  which make  $J_{x^0}(u) < \infty$ , where  $x(t) = x(t, x^0, u)$  and prime indicates transpose. The main result from linear-quadratic optimal control (cf. Sect. 8.4 in [25]) implies that, for AC systems, there is a unique solution  $u$  to this problem, which is given in the following form: there is a matrix  $F \in \mathbb{R}^{m \times n}$  such that solving  $\dot{x} = (A + BF)x$  with  $x(0) = x^0$  gives that  $u(t) := Fx(t)$  minimizes  $J_{x^0}(\cdot)$ . Moreover, this  $F$  stabilizes the system (which is intuitively to be expected, since  $J_{x^0}(u) < \infty$  implies that solutions  $x(t)$  must be in  $L^2$ ), and  $F$  can be computed by the formula

$$F := -R^{-1}B'P, \quad (6)$$

where  $P$  is a symmetric and positive definite solution of the *Matrix Algebraic Riccati Equation*

$$PBR^{-1}B'P - PA - A'P - Q = 0. \quad (7)$$

### A Sufficient Nonlinear Condition

Although of limited applicability, it is worth remarking that there is a partial extension to nonlinear systems of the stabilization method which was just described. For simplicity, we specialize our discussion to control-affine systems, i. e., those for which the input appears only in an affine form. This class is sufficient for the study of most forced mechanical systems. The Eq. (1) becomes:

$$\dot{x} = g_0(x) + \sum_{i=1}^m u_i g_i(x) = g_0(x) + G(x)u.$$



(See Sect. 8.5 in [25] for general  $f$ , and for proofs). We now pick two continuous functions  $Q: \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}$  and  $R: \mathbb{R}^n \rightarrow \mathbb{R}^{n \times n}$ , so that  $R(x)$  is a symmetric positive definite matrix for each  $x$ .

In general, we say that a continuous function

$$V: \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}$$

is *positive definite* if  $V(x) = 0$  only if  $x = 0$ , and it is *proper* (or “weakly coercive”) if for each  $a \geq 0$  the set  $\{x | V(x) \leq a\}$  is compact, or, equivalently,  $V(x) \rightarrow \infty$  as  $|x| \rightarrow \infty$  (radial unboundedness). Given any such  $V$  which is also differentiable, we denote the vector function whose components are the directional derivatives of  $V$  in the directions of the various control vector fields  $g_i$ ,  $i \geq 1$  by:

$$L_G V(x) := \nabla V(x) G(x) = (L_{g_1} V(x), \dots, L_{g_m} V(x)) ,$$

and also write  $L_{g_0} V(x) := \nabla V(x) g_0(x)$ .

We consider the following PDE on such functions  $V$ :

$$\forall x \quad Q(x) + L_{g_0} V(x) - \frac{1}{4} L_G V(x) R(x)^{-1} (L_G V(x))' = 0. \quad (8)$$

This reduces to the Algebraic Riccati Eq. (7) in the special case of linear systems, quadratic  $V(x) = x' P x$  and  $Q(x) = x' Q x$ , and constant matrices  $R(x) \equiv R$ . We also take the following generalization of the feedback law (6):

$$k(x) := -\frac{1}{2} R(x)^{-1} (L_G V(x))' . \quad (9)$$

Finally, we assume that  $Q$  is a positive definite function. One then has [25]:

**Theorem 2** *Suppose that  $V$  is a twice continuously differentiable, positive definite, and proper solution of the PDE (8). Then,  $k$  defined by (9) stabilizes the system.*

This theorem arises from the following optimization problem: for each state  $x^0 \in \mathbb{R}^n$ , minimize the cost

$$J_{x^0}(u) := \int_0^\infty u(t)' R(x(t)) u(t) + Q(x(t)) dt ,$$

where  $x(t) = x(t, x^0, u)$ , over all those controls  $u: [0, \infty) \rightarrow \mathcal{U}$  for which the solution  $x(t, x^0, u)$  of (1) is defined for all  $t \geq 0$  and satisfies  $\lim_{t \rightarrow \infty} x(t) = 0$ . Under the above assumptions, and as for linear systems, one also concludes that for each state  $x^0$  the solution of  $\dot{x} = f(x, k(x))$  with initial state  $x(0) = x^0$  exists for all  $t \geq 0$ , the control  $u(t) = k(x(t))$  is optimal, and  $V(x^0)$  is the optimal cost from initial state  $x^0$ . Moreover, the formula for  $k$  arises

from the Hamilton–Jacobi–Bellman equation of optimal control theory, because

$$k(x) = \operatorname{argmin}_u \{ \nabla V(x) \cdot f(x, u) + u' R(x) u + Q(x) \}$$

when  $f(x, u) = g_0(x) + G(x)u$ .

There are applications where this method has proven useful. Unfortunately, however, and in contrast to the linear case, in general there exists no positive definite, proper, and  $C^2$  solution  $V$  of the above PDE. On the other hand, the formula (9) does appear, with variations, in other contexts, including generalizations of the idea of adding damping to systems, cf. Sect. 5.9 in [25], and, more generally, the use of auxiliary positive definite and proper functions  $V$ , in similar roles, will be central to the control–Lyapunov ideas discussed later.

### Nonlinear Systems: Continuous Feedback

One of the central topics which we will address here concerns possibly discontinuous feedback laws  $k$ . Before turning to that subject, however, we study continuous feedback. When dealing with linear systems, linear feedback is natural, and indeed sufficient from a theoretical standpoint, as shown by the results just reviewed. However, for our general study, major technical questions arise in even deciding on just what degree of regularity should be imposed on the feedback maps  $k$ .

It turns out that the precise requirements away from 0, say asking whether  $k$  is merely continuous or smooth, are not very critical; it is often the case that one can “smooth out” a continuous feedback (or, even, make it real-analytic, via Grauert’s Theorem) away from the origin. So, in order to avoid unnecessary complications in exposition due to nonuniqueness, let us call a feedback  $k$  *regular* if it is locally Lipschitz on  $\mathbb{R}^n \setminus \{0\}$ . For such  $k$ , solutions of initial value problems  $\dot{x} = f(x, k(x))$ ,  $x(0) = x^0$ , are well defined (at least for small time intervals  $[0, \varepsilon)$  and, provided  $k$  is a stabilizing feedback, are unique (cf. Exercise 5.9.9 in [25]).

On the other hand, behavior at the origin cannot be “smoothed out” and, at zero, the precise degree of smoothness plays a central role in the theory [12]. For instance, consider the system

$$\dot{x} = x + u^3 .$$

The continuous (and, in fact, smooth away from zero) feedback  $u = k(x) := -\sqrt[3]{2x}$  globally stabilizes the system (the closed-loop system becomes  $\dot{x} = -x$ ). However, there is no possible stabilizing feedback which is differentiable at the origin, since  $u = k(x) = O(x)$  implies that

$$\dot{x} = x + O(x^3)$$

about  $x = 0$ , which means that the solution starting at any positive and small point moves to the right, instead of towards the origin. (A general result, assuming that  $A$  has no purely imaginary eigenvalues, cf. [25], Section 5.8, is that if – and only if –  $\dot{x} = Ax + Bu + o(x, u)$  can be locally asymptotically stabilized using a feedback which is differentiable at the origin, the linearization  $\dot{x} = Ax + Bu$  must be AC itself. In the example that we gave, this linearization is just  $\dot{x} = x$ , which is not AC.)

We now turn to the question of existence of regular feedback stabilizers. We first study a comparatively trivial case, namely systems with one state variable and one input. After that, we turn to multidimensional systems.

### The Special Case $n = m = 1$

There are algebraic obstructions to the stabilization of  $\dot{x} = f(x, u)$  if the input  $u$  appears nonlinearly in  $f$ . Ignoring the requirement that there be a  $\sigma \in \mathcal{N}$  so that controls can be picked with  $\|u\| \leq \sigma(|x^0|)$ , asymptotic controllability is, for  $n = m = 1$ , equivalent to:

$$(\forall x \neq 0)(\exists u) \ x f(x, u) < 0 \quad (10)$$

(this is proved in [28]; it is fairly obvious, but some care must be taken to deal with the fact that one is allowing arbitrary measurable controls; the argument proceeds by first approximating such controls by piecewise constant ones). Let us introduce the following set:

$$\mathcal{O} := \{(x, u) | x f(x, u) < 0\},$$

and let  $\pi : (x, u) \mapsto x$  be the projection into the first coordinate in  $\mathbb{R}$ . Then, (10) is equivalent to:

$$\pi \mathcal{O} = \mathbb{R} \setminus \{0\}.$$

(One can easily include the requirement “ $\|u\| \leq \sigma(|x^0|)$ ” by asking that for each interval  $[-K, K] \subset \mathbb{R}$  there must be some compact set  $C_K \subset \mathbb{R}^2$  so that  $[-K, K] \subseteq \pi(C_K)$ . For simplicity, we ignore this technicality.) In these terms, a stabilizing feedback is nothing else than a locally bounded map  $k : \mathbb{R} \rightarrow \mathbb{R}$  such that  $k(0) = 0$  and so that  $k$  is a section of  $\pi$  on  $\mathbb{R} \setminus \{0\}$ :

$$(x, k(x)) \in \mathcal{O} \quad \forall x \neq 0.$$

For a regular feedback, we ask that  $k$  be locally Lipschitz on  $\mathbb{R} \setminus \{0\}$ .

Clearly, there is no reason for Lipschitz, or for that matter, just continuous, sections of  $\pi$  to exist. As an illustration, take the system

$$\dot{x} = x [(u - 1)^2 - (x - 1)] [(u + 1)^2 + (x - 2)].$$

Let

$$\mathcal{O}_1 = \{(u + 1)^2 < (2 - x)\} \text{ and } \mathcal{O}_2 = \{(u - 1)^2 < (x - 1)\}$$

(these are the interiors of two disjoint parabolas). Here,  $\mathcal{O}$  has three connected components, namely  $\mathcal{O}_2$ ,  $\mathcal{O}_1$  intersected with  $x < 0$ , and  $\mathcal{O}_1$  intersected with  $x < 0$ . It is clear that, even though  $\pi \mathcal{O} = \mathbb{R}$ , there is no continuous curve (graph of  $u = k(x)$ ) which is always in  $\mathcal{O}$  and projects onto  $\mathbb{R} \setminus \{0\}$ . On the other hand, there exist many possible feedback stabilizers provided that we allow one discontinuity. It is also possible to provide examples, even with  $f(x, u)$  smooth, for which an infinite number of switches are needed in any possible stabilizing feedback law. Finally, it may even be possible to stabilize semiglobally with a regular feedback, meaning that for each compact subset  $K$  of the state-space there is a regular, even smooth, feedback law  $u = k_K(x)$  such that all states in  $K$  get driven asymptotically to the origin, but yet it may be impossible to find a single  $u = k(x)$  which works globally. See [26] for details.

When feedback laws are required to be continuous at the origin, new obstructions arise. The case of systems with  $n = m = 1$  is also a good way to introduce this subject. The first observation is that stabilization about the origin (even if just local) means that we must have, near zero:

$$f(x, k(x)) \begin{cases} > 0 & \text{if } x < 0 \\ < 0 & \text{if } x > 0 \\ = 0 & \text{if } x = 0 \end{cases}.$$

In fact, all that we need is that  $f(x_1, k(x_1)) < 0$  for some  $x_1 > 0$  and  $f(x_2, k(x_2)) > 0$  for some  $x_2 < 0$ . This guarantees, via the intermediate-value theorem that, if  $k$  is continuous, the projection

$$(-\varepsilon, \varepsilon) \rightarrow \mathbb{R}, \quad x \mapsto f(x, k(x))$$

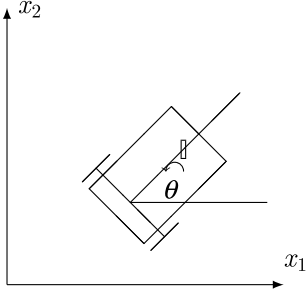
is onto a neighborhood of zero, for each  $\varepsilon > 0$ . It follows, in particular, that the image of

$$(-\varepsilon, \varepsilon) \times (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}, \quad (x, u) \mapsto f(x, u)$$

also contains a neighborhood of zero, for any  $\varepsilon > 0$  (that is, the map  $(x, u) \mapsto f(x, u)$  is open at zero). This last property is intrinsic, being stated in terms of the original data  $f(x, u)$  and not depending upon the feedback  $k$ . Brockett's condition, to be described next, is a far-reaching generalization of this argument; in its proof, degree theory replaces the use of the intermediate value theorem.

### Obstructions and Necessary Degree Conditions

If there are “obstacles” in the state-space, or more precisely if the state-space is a proper subset of  $\mathbb{R}^n$ , discontinuities in feedback laws cannot in general be avoided,



Stability and Feedback Stabilization, Figure 1  
Shopping cart

since the domain of attraction of an asymptotically stable vector field must be diffeomorphic to Euclidean space. But even if states evolve in Euclidean spaces, similar obstructions may arise. These are due not to the topology of the state space, but to “virtual obstacles” implicit in the form of the system equations. These obstacles occur when it is impossible to move instantaneously in certain directions, even if it is possible to move eventually in every direction, the phenomenon of “nonholonomy”. As an illustration, let us consider a model for the “shopping cart” shown in Fig. 1 (“knife-edge” or “unicycle” are other names for this example). The state is given by the orientation  $\theta$ , together with the coordinates  $x_1, x_2$  of the midpoint between the back wheels. The front wheel is a castor, free to rotate. There is a non-slipping constraint on movement: the velocity  $(\dot{x}_1, \dot{x}_2)'$  must be parallel to the vector  $(\cos \theta, \sin \theta)'$ . This leads to the following equations:

$$\begin{aligned}\dot{x}_1 &= u_1 \cos \theta \\ \dot{x}_2 &= u_1 \sin \theta \\ \dot{\theta} &= u_2\end{aligned}$$

where we may view  $u_1$  as a “drive” command and  $u_2$  as a steering control. (In practice, one would implement these controls by means of differential forces on the two back corners of the cart.) The feedback transformation  $z_1 := \theta$ ,  $z_2 := x_1 \cos \theta + x_2 \sin \theta$ ,  $z_3 := x_1 \sin \theta - x_2 \cos \theta$ ,  $v_1 := u_2$ , and  $v_2 := u_1 - u_2 z_3$  brings the system into the system with equations  $\dot{z}_1 = v_1$ ,  $\dot{z}_2 = v_2$ ,  $\dot{z}_3 = z_2 v_1$  known as “Brockett’s example” or “nonholonomic integrator” (yet another change can bring the third equation into the form  $\dot{z}_3 = z_1 v_2 - z_2 v_1$ ). We view the system as having state space  $\mathbb{R}^3$ . Although a physically more accurate state space would be the manifold  $\mathbb{R}^2 \times \mathbb{S}^1$ , the necessary condition to be given is of a local nature, so the global structure is unimportant.

This system is (obviously) completely controllable (formally, controllability can be checked using the Lie al-

gebra rank condition, as in e. g. [25], Exercise 4.3.16), and in particular is AC. However, we may expect that discontinuities are unavoidable due to the non-slip constraint, which does not allow moving from, for example the position  $x_1 = 0$ ,  $\theta = 0$ ,  $x_2 = 1$  in a straight line towards the origin. Indeed, one then has [3]:

**Theorem 3** *If there is a stabilizing feedback which is regular and continuous at zero, then the map  $(x, u) \mapsto f(x, u)$  is open at zero.*

The test fails here, since no points of the form  $(0, \varepsilon, *)$  belong to the image of the map

$$\begin{aligned}\mathbb{R}^5 \rightarrow \mathbb{R}^3: (x_1, x_2, \theta, u_1, u_2)' &\mapsto f(x, u) \\ &= (u_1 \cos \theta, u_1 \sin \theta, u_2)'\end{aligned}$$

for  $\theta \in (-\pi/2, \pi/2)$ , unless  $\varepsilon = 0$ .

More generally, it is impossible to continuously stabilize any system without drift

$$\dot{x} = u_1 g_1(x) + \dots + u_m g_m(x) = G(x)u$$

if  $m < n$  and  $\text{rank}[g_1(0), \dots, g_m(0)] = m$  (this includes all totally nonholonomic mechanical systems). Indeed, under these conditions, the map  $(x, u) \mapsto G(x)u$  cannot contain a neighborhood of zero in its image, when restricted to a small enough neighborhood of zero. Indeed, let us first rearrange the rows of  $G$ :

$$G(x) \rightsquigarrow \begin{pmatrix} G_1(x) \\ G_2(x) \end{pmatrix}$$

so that  $G_1(x)$  is of size  $m \times m$  and is nonsingular for all states  $x$  that belong to some neighborhood  $N$  of the origin. Then,

$$\begin{pmatrix} 0 \\ a \end{pmatrix} \in \text{Im}[N \times \mathbb{R}^m \rightarrow \mathbb{R}^n: (x, u) \mapsto G(x)u] \Rightarrow a = 0$$

(since  $G_1(x)u = 0 \Rightarrow u = 0 \Rightarrow G_2(x)u = 0$  too).

If the condition  $\text{rank}[g_1(0), \dots, g_m(0)] = m$  is violated, we cannot conclude a negative result. For instance, the system  $\dot{x}_1 = x_1 u$ ,  $\dot{x}_2 = x_2 u$  has  $m = 1 < 2 = n$  but it can be stabilized by means of the feedback law  $u = -(x_1^2 + x_2^2)$ .

Observe that for linear systems, Brockett’s condition says that

$$\text{rank}[A, B] = n$$

which is the Hautus controllability condition (see e. g. [25], Lemma 3.3.7) at the zero mode.

**Idea of the Proof** One may prove Brockett's condition in several ways. A proof based on degree theory is probably easiest, and proceeds as follows (for details see for instance [25], Sect 5.9). The basic fact, due to Krasnosel'ski, is that if the system  $\dot{x} = F(x) = f(x, k(x))$  has the origin as an asymptotically stable point and  $F$  is regular (since  $k$  is), then the degree (index) of  $F$  with respect to zero is  $(-1)^n$ , where  $n$  is the system dimension. In particular, the degree is also nonzero with respect to points  $p$  in a neighborhood of 0, which means that the equation  $F(x) = p$  can be solved for small  $p$ , and hence  $f(x, u) = p$  can be solved as well. The proof that the degree is  $(-1)^n$  follows by exhibiting a homotopy, namely

$$F_t(x^0) = \frac{1}{t} \left[ x \left( \frac{t}{1-t}, x^0 \right) - x^0 \right],$$

between  $F_0 = F$  and  $F_1(x) = -x$ , and noting that the degree of the latter is obviously  $(-1)^n$ . An alternative proof uses Lyapunov functions. Asymptotic stability implies the existence of a smooth Lyapunov function  $V$  for  $\dot{x} = F(x) = f(x, k(x))$ , so, on the boundary  $\partial B$  of a sublevel set  $B = \{x | V(x) \leq c\}$  we have that  $F$  points towards the interior of  $B$ . Thus, for  $p$  small,  $F(x) - p$  still points to the interior, which means that  $B$  is invariant with respect to the perturbed vector field  $\dot{x} = F(x) - p$ . Provided that a fixed-point theorem applies to continuous maps  $B \rightarrow B$ , this implies that  $F(x) - p$  must vanish somewhere in  $B$ , that is, the equation  $F(x) = p$  can be solved. (Because, for each small  $h > 0$ , the time- $h$  flow  $\phi$  of  $F - p$  has a fixed point  $x_h \in B$ , i.e.  $\phi(h, x_h) = x_h$ , so picking a convergent subsequence  $x_h \rightarrow \bar{x}$  gives that  $0 = (\phi(h, x_h) - x_h)/(h) \rightarrow F(\bar{x}) - p$ .) A fixed point theorem can indeed be applied, because  $B$  is a retract of  $\mathbb{R}^n$  (use the flow itself); note that this argument gives a weaker conclusion than the degree condition.

### Control-Lyapunov Functions

The method of control-Lyapunov functions ("clf's") provides a powerful tool for studying stabilization problems, both as a basis of theoretical developments and as a method for actual feedback design.

Before discussing clfs, let us quickly review the classical concept of Lyapunov functions, through a simple example. Consider first a damped spring-mass system  $\ddot{y} + \dot{y} + y = 0$ , or, in state-space form with  $x_1 = y$  and  $x_2 = \dot{y}$ ,  $\dot{x}_1 = x_2$ ,  $\dot{x}_2 = -x_1 - x_2$ . One way to verify global asymptotic stability of the equilibrium  $x = 0$  is to pick the (Lyapunov) function  $V(x_1, x_2) := \frac{3}{2}x_1^2 + x_1x_2 + x_2^2$ , and observe that  $\nabla V(x) \cdot f(x) = -|x|^2 < 0$  if  $x \neq 0$ , which

means that

$$\frac{dV(x(t))}{dt} = -|x(t)|^2 < 0$$

along all nonzero solutions, and thus the energy-like function  $V$  decreases along all trajectories, which, since  $V$  is a nondegenerate quadratic form, implies that  $x(t)$  decreases, and in fact  $x(t) \rightarrow 0$ . Of course, in this case one could compute solutions explicitly, or simply note that the characteristic equation has all roots with negative real part, but Lyapunov functions are a general technique. (In fact, the classical converse theorems of Massera and Kurzweil [17,19] show that, whenever a system is GAS, there always exists a smooth Lyapunov function  $V$ .)

Now let us modify this example to deal with a control system, and consider a forced (but undamped) harmonic oscillator  $\ddot{x} + x = u$ , i.e.  $\dot{x}_1 = x_2$ ,  $\dot{x}_2 = -x_1 + u$ . The damping feedback  $u = -x_2$  stabilizes the system, but let us pretend that we do not know that. If we take the same  $V$  as before, now the derivatives along trajectories are, using " $\dot{V}(x, u)$ " to denote  $\nabla V(x) \cdot f(x, u)$  and omitting arguments  $t$  in  $x(t)$  and  $u(t)$ :

$$\dot{V}(x, u) = -x_1^2 + x_1x_2 + x_2^2 - (x_1 + 2x_2)u.$$

This expression is affine in  $u$ . Thus, if  $x$  is a state such that  $x_1 + 2x_2 \neq 0$ , then we may pick a control value  $u$  (which depends on this current state  $x$ ) such that  $\dot{V} < 0$ . On the other hand, if  $x_1 + 2x_2 = 0$ , then the expression reduces to  $\dot{V} = -5x_2^2$  (for any  $u$ ), which is negative unless  $x_2$  (and hence also  $x_1 = -2x_2$ ) vanishes.

In conclusion, for each  $x \neq 0$  there is some  $u$  so that  $\dot{V}(x, u) < 0$ . This is, except for some technicalities to be discussed, the characterizing property of control-Lyapunov functions. For any given compact subset  $B$  in  $\mathbb{R}^n$ , we now pick some compact subset  $\mathcal{U}_0 \subset \mathcal{U}$  so that

$$\forall x \in B, x \neq 0, \quad \exists u \in \mathcal{U}_0 \quad \text{such that} \quad \dot{V}(x, u) < 0. \quad (11)$$

In principle, then, we could then stabilize the system, for states in  $B$ , by using the steepest descent feedback law:

$$k(x) := \underset{u \in \mathcal{U}_0}{\operatorname{argmin}} \nabla V(x) \cdot f(x, u) \quad (12)$$

("argmin" means "pick any  $u$  at which the min is attained"; we restricted  $\mathcal{U}$  to be assured that  $\dot{V}(x, u)$  attains a minimum). Note that the stabilization problem becomes, in these terms, a set of static nonlinear programming problems: minimize a function of  $u$ , for each  $x$ . Global stabilization is also possible, by appropriately picking  $\mathcal{U}_0$  as a func-

tion of the norm of  $x$ ; later we discuss a precise formulation.

Control–Lyapunov functions, if understood non-technically as the basic paradigm “look for a function  $V(x)$  with the properties that  $V(x) \approx 0$  if and only if  $x \approx 0$ , and so that for each  $x \neq 0$  it is possible to decrease  $V(x)$  by some control action,” constitute a very general approach to control (sometimes expressed in a dual fashion, as maximization of some measure of success). They appear in such disparate areas as A.I. game-playing programs (position evaluations), energy arguments for dissipative systems, program termination (Floyd/Dijkstra “variant”), and learning control (“critics” implemented by neural networks). More relevantly to this paper, the idea underlies much of modern feedback control design, as illustrated for instance by the books [7,11,15,16,25].

**Differentiable clfs: Precise Definition** A *differentiable control–Lyapunov function* (clf) is a differentiable function  $V: \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}$  which is proper, positive definite, and *infinitesimally decreasing*, meaning that there exists a positive definite continuous function  $W: \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}$ , and there is some  $\sigma \in \mathcal{N}$ , so that

$$\sup_{x \in \mathbb{R}^n} \min_{|u| \leq \sigma(|x|)} \nabla V(x) \cdot f(x, u) + W(x) \leq 0. \quad (13)$$

This is basically the same as condition (11), with  $\mathcal{U}_0$  = the ball of radius  $\sigma(|x|)$  picked as a function of  $x$ . The main difference is that, instead of saying “ $\nabla V(x) \cdot f(x, u) < 0$  for  $x \neq 0$ ” we write  $\nabla V(x) \cdot f(x, u) \leq -W(x)$ , where  $W$  is negative when  $x \neq 0$ . The two definitions are equivalent, but the “Hamiltonian” version used here is the correct one for the generalizations to be given, to nonsmooth  $V$ .

The basic result is due to Artstein [2]:

**Theorem 4** A control-affine system  $\dot{x} = g_0(x) + \sum u_i g_i(x)$  admits a differentiable clf if and only if it admits a regular stabilizing feedback.

The proof of sufficiency is easy: if there is such a  $k$ , then the converse Lyapunov theorem, applied to the closed-loop system  $F(x) = f(x, k(x))$ , provides a smooth  $V$  such that

$$L_F V(x) = \nabla V(x) F(x) < 0 \quad \forall x \neq 0.$$

This gives that for all nonzero  $x$  there is some  $u$  (bounded on bounded sets, because  $k$  is locally bounded by definition of feedback) so that  $\dot{V}(x, u) < 0$ ; and one can put this in the form (13).

The necessity is more interesting. The original proof in [2] proceeds by a nonconstructive argument involving partitions of unity, but it is also possible [24,25] to exhibit

explicitly a feedback, written as a function:

$$k(\nabla V(x) \cdot g_0(x), \dots, \nabla V(x) \cdot g_m(x))$$

of the directional derivatives of  $V$  along the vector fields defining the system (*universal formulas* for stabilization). Taking for simplicity  $m = 1$ , one such formula is:

$$k(x) := -\frac{a(x) + \sqrt{a(x)^2 + b(x)^4}}{b(x)} \quad (0 \text{ if } b = 0)$$

where  $a(x) := \nabla V(x) \cdot g_0(x)$  and  $b(x) := \nabla V(x) \cdot g_1(x)$ . The expression for  $k$  is analytic in  $a, b$  when  $x \neq 0$ , because the clf property means that  $a(x) < 0$  whenever  $b(x) = 0$  [24,25].

Thus, the question of existence of regular feedback, for control-affine systems, reduces to the search for differentiable clfs, and this gives rise to a vast literature dealing with the construction of such  $V$ 's, see [7,15,16,25] and references therein. Many other theoretical issues are also answered by Artstein's theorem. For example, via Kurzweil's converse theorem one has that the existence of  $k$  merely continuous on  $\mathbb{R}^n \setminus \{0\}$  suffices for the existence of smooth (infinitely differentiable)  $V$ , and from here one may in turn find a  $k$  which is smooth on  $\mathbb{R}^n \setminus \{0\}$ . In addition, one may easily characterize the existence of  $k$  continuous at zero as well as regular: this is equivalent to the *small control property*: for each  $\varepsilon > 0$  there is some  $\delta > 0$  so that  $0 < |x| < \delta$  implies that  $\min_{|u| \leq \varepsilon} \nabla V(x) \cdot f(x, u) < 0$  (if this property holds, the universal formula automatically provides such a  $k$ ). We should note that Artstein provided a result valid for general, not necessarily control-affine systems  $\dot{x} = f(x, u)$ ; however, the obtained “feedback” has values in sets of relaxed controls, and is not a feedback law in the classical sense. Later, we discuss a different generalization.

Differentiable clfs will in general not exist, because of obstructions to regular feedback stabilization. This leads us naturally into the twin subjects of discontinuous feedbacks and non-differentiable clfs.

## Discontinuous Feedback

The previous results and examples show that, in order to develop a satisfactory general theory of stabilization, one in which one proves the implication “asymptotic controllability implies feedback stabilizability,” we must allow discontinuous feedback laws  $u = k(x)$ . But then, a major technical difficulty arises: solutions of the initial-value problem  $\dot{x} = f(x, k(x))$ ,  $x(0) = x^0$ , interpreted in the classical sense of differentiable functions or even as (absolutely) continuous solutions of the integral equation



tion  $x(t) = x^0 + \int_0^t f(x(s), k(x(s)))ds$ , do not exist in general. The only general theorems apply to systems  $\dot{x} = F(x)$  with continuous  $F$ . For example, there is no solution to  $\dot{x} = -\text{sign } x$ ,  $x(0) = 0$ , where  $\text{sign } x = -1$  for  $x < 0$  and  $\text{sign } x = 1$  for  $x \geq 0$ . So one cannot even pose the stabilization problem in a mathematically consistent sense.

There is, of course, an extensive literature addressing the question of discontinuous feedback laws for control systems and, more generally, differential equations with discontinuous right-hand sides. One of the best-known candidates for the concept of solution of (3) is that of a *Filippov solution* [6,9], which is defined as the solution of a certain differential inclusion with a multivalued right-hand side which is built from  $f(x, k(x))$ . Unfortunately, there is no hope of obtaining the implication “asymptotic controllability implies feedback stabilizability” if one interprets solutions of (3) as Filippov solutions. This is a consequence of results in [5,22], which established that the existence of a discontinuous stabilizing feedback in the Filippov sense implies the Brockett necessary conditions, and, moreover, for systems affine in controls it also implies the existence of regular feedback (which we know is in general impossible).

A different concept of solution originates with the theory of discontinuous positional control developed by Krasovskii and Subbotin in the context of differential games in [14], and it is the basis of the new approach to discontinuous stabilization proposed in [4], to which we now turn.

### Limits of High-Frequency Sampling

By a *sampling schedule* or *partition*  $\pi = \{t_i\}_{i \geq 0}$  of  $0, +\infty$  we mean an infinite sequence

$$0 = t_0 < t_1 < t_2 < \dots$$

with  $\lim_{i \rightarrow \infty} t_i = \infty$ . We call

$$\mathbf{d}(\pi) := \sup_{i \geq 0} (t_{i+1} - t_i)$$

the *diameter* of  $\pi$ . Suppose that  $k$  is a given feedback law for system (1). For each  $\pi$ , the  $\pi$ -trajectory starting from  $x^0$  of system (3) is defined recursively on the intervals  $[t_i, t_{i+1})$ ,  $i = 0, 1, \dots$ , as follows. On each interval  $t_i, t_{i+1})$ , the initial state is measured, the control value  $u_i = k(x(t_i))$  is computed, and the constant control  $u \equiv u_i$  is applied until time  $t_{i+1}$ ; the process is then iterated. That is, we start with  $x(t_0) = x^0$  and solve recursively

$$\dot{x}(t) = f(x(t), k(x(t_i))), \quad t \in [t_i, t_{i+1}), \quad i = 0, 1, 2, \dots$$

using as initial value  $x(t_i)$  the endpoint of the solution on the preceding interval. The ensuing  $\pi$ -trajectory, which we denote as  $x_\pi(\cdot, x^0)$ , is defined on some maximal nontrivial interval; it may fail to exist on the entire interval  $[0, +\infty)$  due to a blow-up on one of the subintervals  $t_i, t_{i+1})$ . We say that it is *well defined* if  $x_\pi(t, x^0)$  is defined on all of  $[0, +\infty)$ .

**Definition** The feedback  $k: \mathbb{R}^n \rightarrow \mathcal{U}$  stabilizes the system (1) if there exists a function  $\beta \in \mathcal{KL}$  so that the following property holds: For each

$$0 < \varepsilon < K$$

there exists a  $\delta = \delta(\varepsilon, K) > 0$  such that, for every sampling schedule  $\pi$  with  $\mathbf{d}(\pi) < \delta$ , and for each initial state  $x^0$  with  $|x^0| \leq K$ , the corresponding  $\pi$ -trajectory of (3) is well-defined and satisfies

$$|x_\pi(t, x^0)| \leq \max\{\beta(K, t), \varepsilon\} \quad \forall t \geq 0. \quad (14)$$

In particular, we have

$$|x_\pi(t, x^0)| \leq \max\{\beta(|x^0|, t), \varepsilon\} \quad \forall t \geq 0 \quad (15)$$

whenever  $0 < \varepsilon < |x^0|$  and  $\mathbf{d}(\pi) < \delta(\varepsilon, |x^0|)$  (just take  $K := |x^0|$ ).

Observe that the role of  $\delta$  is to specify a lower bound on intersampling times. Roughly, one is requiring that

$$t_{i+1} \leq t_i + \theta(|x(t_i)|)$$

for each  $i$ , where  $\theta$  is an appropriate positive function.

Our definition of stabilization is physically meaningful, and is very natural in the context of sampled-data (computer control) systems. It says in essence that a feedback  $k$  stabilizes the system if it drives all states asymptotically to the origin and with small overshoot when using *any fast enough sampling schedule*. A high enough sampling frequency is generally required when close to the origin, in order to guarantee small displacements, and also at infinity, so as to preclude large excursions or even blow-ups in finite time. This is the reason for making  $\delta$  depend on  $\varepsilon$  and  $K$ .

This concept of stabilization can be reinterpreted in various ways. One is as follows. Pick any initial state  $x^0$ , and consider any sequence of sampling schedules  $\pi_\ell$  whose diameters  $\mathbf{d}(\pi_\ell)$  converge to zero as  $\ell \rightarrow \infty$  (for instance, constant sampling rates with  $t_i = i/\ell$ ,  $i = 0, 1, 2, \dots$ ). Note that the functions  $x_\ell := x_{\pi_\ell}(\cdot, x^0)$  remain in a bounded set, namely the ball of radius  $\beta(|x^0|, 0)$  (at least for  $\ell$  large enough, for instance, any  $\ell$  so that  $\mathbf{d}(\pi_\ell)\delta(|x^0|/2, |x^0|)$ ). Because  $f(x, k(x))$  is bounded

on this ball, these functions are equicontinuous, and (Arzela–Ascoli’s Theorem) we may take a subsequence, which we denote again as  $\{x_\ell\}$ , so that  $x_\ell \rightarrow x$  as  $\ell \rightarrow \infty$  (uniformly on compact time intervals) for some absolutely continuous (even Lipschitz) function  $x: [0, \infty) \rightarrow \mathbb{R}^n$ . We may think of any limit function  $x(\cdot)$  that arises in this fashion as a generalized solution of the closed-loop Eq. (3). That is, generalized solutions are the limits of trajectories arising from arbitrarily high-frequency sampling when using the feedback law  $u = k(x)$ . Generalized solutions, for a given initial state  $x^0$ , may not be unique – just as may happen with continuous but non-Lipschitz feedback – but there is always existence, and, moreover, for any generalized solution,  $|x(t)| \leq \beta(|x^0|, t)$  for all  $t \geq 0$ . This is precisely the defining estimate for the GAS property. Moreover, if  $k$  happens to be regular, then the unique solution of  $\dot{x} = f(x, k(x))$  in the classical sense is also the unique generalized solution, so we have a reasonable extension of the concept of solution. (This type of interpretation is somewhat analogous, at least in spirit, to the way in which “relaxed” controls are interpreted in optimal trajectory calculations, namely through high-frequency switching of approximating regular controls.) The definition of stabilization was given in [4] in a slightly different form; see [26] for a discussion of the equivalence.

### Stabilizing Feedbacks Exist

The main result is [4]:

**Theorem 5** *The system (1) admits a stabilizing feedback if and only if it is asymptotically controllable.*

Necessity is clear. The sufficiency statement is proved by construction of  $k$ , and is based on the following ingredients:

- Existence of a nonsmooth control-Lyapunov function  $V$ .
- Regularization on shells of  $V$ .
- Pointwise minimization of a Hamiltonian for the regularized  $V$ .

In order to sketch this construction, we start by quickly reviewing a basic concept from nonsmooth analysis.

**Proximal Subgradients** Let  $V$  be any continuous function  $\mathbb{R}^n \rightarrow \mathbb{R}$  (or even, just lower semicontinuous and with extended real values). A *proximal subgradient* of  $V$  at the point  $x \in \mathbb{R}^n$  is any vector  $\zeta \in \mathbb{R}^n$  such that, for some  $\sigma > 0$  and some neighborhood  $\mathcal{O}$  of  $x$ ,

$$V(y) \geq V(x) + \zeta \cdot (y - x) - \sigma^2 |y - x|^2 \quad \forall y \in \mathcal{O}.$$

In other words, proximal subgradients are the possible gradients of supporting quadratics at the point  $x$ . The set of all proximal subgradients at  $x$  is denoted  $\partial_p V(x)$ .

**Nonsmooth Control–Lyapunov Functions** A continuous (but not necessarily differentiable)  $V: \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}$  is a *control-Lyapunov function* (clf) if it is proper, positive definite, and infinitesimally decreasing in the following generalized sense: there exist a positive definite continuous  $W: \mathbb{R}^n \rightarrow \mathbb{R}_{\geq 0}$  and a  $\sigma \in \mathcal{N}$  so that

$$\sup_{x \in \mathbb{R}^n} \max_{\zeta \in \partial_p V(x)} \min_{|u| \leq \sigma(|x|)} \zeta \cdot f(x, u) + W(x) \leq 0. \quad (16)$$

This is the obvious generalization of the differentiable case in (13); we are still asking that one should be able to make  $\nabla V(x) \cdot f(x, u) < 0$  by an appropriate choice of  $u = u_x$ , for each  $x \neq 0$ , except that now we replace  $\nabla V(x)$  by the proximal subgradient set  $\partial_p V(x)$ . An equivalent property is to ask that  $V$  be a viscosity supersolution of the corresponding Hamilton–Jacobi–Bellman equation.

For nonsmooth clfs, the main basic result is [23,26]:

**Theorem 6** *The system (1) is asymptotically controllable if and only if it admits a continuous clf.*

The proof is based on first constructing an appropriate  $W$ , and then letting  $V$  be the optimal cost (Bellman function) for the problem  $\min \int_0^\infty W(x(s))ds$ . However, some care has to be taken to insure that  $V$  is continuous, and the cost has to be adjusted in order to deal with possibly unbounded minimizers. An important and very useful refinement of this result is the fact that a locally Lipschitz clf can also be shown to exist [21].

**Regularization** Once  $V$  is known to exist, the next step in the construction of a stabilizing feedback is to obtain Lipschitz approximations of  $V$ . For this purpose, one considers the Iosida–Moreau inf-convolution of  $V$  with a quadratic function:

$$V_\alpha(x) := \inf_{y \in \mathbb{R}^n} \left[ V(y) + \frac{1}{2\alpha^2} |y - x|^2 \right]$$

where the number  $\alpha > 0$  is picked constant on appropriate regions. One has that  $V_\alpha(x) \nearrow V(x)$ , uniformly on compacts. Since  $V_\alpha$  is locally Lipschitz, Rademacher’s Theorem insures that  $V_\alpha$  is differentiable almost everywhere. The feedback  $k$  is then made equal to a pointwise minimizer  $k_\alpha$  of the Hamiltonian, at the points of differentiability (compare with (12) for the case of differentiable  $V$ ):

$$k_\alpha(x) := \operatorname{argmin}_{u \in \mathcal{U}_0} \nabla V_\alpha(x) \cdot f(x, u),$$

where  $\alpha$  and the compact  $\mathcal{U}_0 = \mathcal{U}_0(\alpha)$  are chosen constant on certain compacts and this choice is made in between level curves. The critical fact is that  $V_\alpha$  is itself a clf for the original system, at least when restricted to the region where it is needed. More precisely, on each shell of the form

$$C = \{x \in \mathbb{R}^n \mid r \leq |x| \leq R\},$$

there are positive numbers  $m$  and  $\alpha_0$  and a compact subset  $\mathcal{U}_0$  such that, for each  $0 < \alpha \leq \alpha_0$ , each  $x \in C$ , and every  $\zeta \in \partial_p V_\alpha(x)$ ,

$$\min_{u \in \mathcal{U}_0} \zeta \cdot f(x, u) + m \leq 0.$$

See [26]. (Actually, this description is oversimplified, and the proof is a bit more delicate. One must define, on appropriate compact sets

$$k(x) := \operatorname{argmin}_{u \in \mathcal{U}_0} \zeta_\alpha(x) \cdot f(x, u),$$

where  $\zeta_\alpha(x)$  is carefully chosen. At points  $x$  of nondifferentiability,  $\zeta_\alpha(x)$  is not a proximal subgradient of  $V_\alpha$ , since  $\partial_p V_\alpha(x)$  may well be empty. One uses, instead, the fact that  $\zeta_\alpha(x)$  happens to be in  $\partial_p V(x')$  for some  $x' \approx x$ . See [4] for details.)

### Sensitivity to Small Measurement Errors

We have seen that every asymptotically controllable system admits a feedback stabilizer  $k$ , generally discontinuous, which renders the closed-loop system  $\dot{x} = f(x, k(x))$  GAS. On the other hand, one of the main motivations for the use of feedback is in order to deal with uncertainty, and one possible source of uncertainty are measurement errors in state estimation. The use of discontinuous feedback means that undesirable behavior – chattering – may arise. In fact, one of the main reasons for the focus on continuous feedback is precisely in order to avoid such behaviors. Thus, we turn now to an analysis of the effect of measurement errors. Suppose first that  $k$  is a continuous function of  $x$ . Then, if the error  $e$  is small, using the control  $u' = k(x + e)$  instead of  $u = k(x)$  results in behavior which remains close to the intended one, since  $k(x + e) \approx k(x)$ ; moreover, if  $e \ll x$  then stability is preserved. This property of robustness to small errors when  $k$  is continuous can be rigorously established by means of a Lyapunov proof, based on the observation that, if  $V$  is a Lyapunov function for the closed-loop system, then continuity of  $f(x, k(x + e))$  on  $e$  means that

$$\nabla V(x) \cdot f(x, k(x + e)) \approx \nabla V(x) \cdot f(x, k(x)) < 0.$$

Unfortunately, when  $k$  is not continuous, this argument breaks down. However, it can be modified so as to avoid invoking continuity of  $k$ . Assuming that  $V$  is continuously differentiable, one can argue that

$$\nabla V(x) \cdot f(x, k(x + e)) \approx \nabla V(x + e) \cdot f(x, k(x + e)) < 0$$

(using the Lyapunov property at the point  $x + e$  instead of at  $x$ ). This observation leads to a theorem, formulated below, which says that a discontinuous feedback stabilizer, robust with respect to small observation errors, can be found provided that there is a  $C^1$  clf. In general, as there are no  $C^1$ , but only continuous, clf's, one may not be able to find any feedback law that is robust in this sense.

There are many well-known techniques for avoiding chattering, and a very common one is the introduction of deadzones where no action is taken. The feedback constructed in [4], with no modifications needed, can always be used in a manner robust with respect to small observation errors, using such an approach. Roughly speaking, the general idea is as follows. Suppose that the true current state, let us say at time  $t = t_i$ , is  $x$ , but that the controller uses  $u = k(\tilde{x})$ , where  $\tilde{x} = x + e$ , and  $e$  is small. Call  $x'$  the state that results at the next sampling time,  $t = t_{i+1}$ . By continuity of solutions on initial conditions,  $|x' - \tilde{x}'|$  is also small, where  $\tilde{x}'$  is the state that would have resulted from applying the control  $u$  if the true state had been  $\tilde{x}$ . By continuity, it follows that  $V_\alpha(x) \approx V_\alpha(\tilde{x})$  and also  $V_\alpha(x') \approx V_\alpha(\tilde{x}')$ . On the other hand, the construction in [4] provides that  $V_\alpha(\tilde{x}') < V_\alpha(\tilde{x}) - d(t_{i+1} - t_i)$ , where  $d$  is some positive constant (this is valid while we are far from the origin). Hence, if  $e$  is sufficiently small compared to the intersample time  $t_{i+1} - t_i$ , it will necessarily be the case that  $V_\alpha(x')$  must also be smaller than  $V_\alpha(x)$ . This discussion may be formalized in several ways; see [26] for a precise statement.

If we insist upon fast sampling, a necessary condition arises, as was proved in the recent paper [18] (which, in turn, represented an extension of the work by Hermes [10] for classical solutions under observation error). We next discuss the main result from that paper. We consider systems

$$\dot{x}(t) = f(x(t), k(x(t) + e(t)) + d(t)) \quad (17)$$

in which there are observation errors as well as, now, possible actuator errors  $d(\cdot)$ . Actuator errors  $d(\cdot): [0, \infty) \rightarrow \mathcal{U}$  are Lebesgue measurable and locally essentially bounded, and observation errors  $e(\cdot): [0, \infty) \rightarrow \mathbb{R}^n$  are locally bounded. We define solutions of (17), for each sampling schedule  $\pi$ , in the usual manner, i.e., solving recursively on the intervals  $t_i, t_{i+1}$ ,  $i = 0, 1, \dots$ , the differen-

tial equation

$$\dot{x}(t) = f(x(t), k(x(t_i) + e(t_i)) + d(t)) \quad (18)$$

with  $x(0) = x^0$ . We write  $x(t) = x_\pi(t, x^0, d, e)$  for the solution, and say it is *well-defined* if it is defined for all  $t \geq 0$ .

**Definition** The feedback  $k: \mathbb{R}^n \rightarrow \mathcal{U}$  stabilizes the system (17) if there exists a function  $\beta \in \mathcal{KL}$  so that the following property holds: For each

$$0 < \varepsilon < K$$

there exist  $\delta = \delta(\varepsilon, K) > 0$  and  $\eta = \eta(\varepsilon, K)$  such that, for every sampling schedule  $\pi$  with  $d(\pi) < \delta$ , each initial state  $x^0$  with  $|x^0| \leq K$ , and each  $e, d$  such that  $|e(t)| \leq \eta$  for all  $t \geq 0$  and  $|d(t)| \leq \eta$  for almost all  $t \geq 0$ , the corresponding  $\pi$ -trajectory of (17) is well-defined and satisfies

$$|x_\pi(t, x^0, d, e)| \leq \max\{\beta(K, t), \varepsilon\} \quad \forall t \geq 0. \quad (19)$$

In particular, taking  $K := |x^0|$ , one has that

$$|x_\pi(t, x^0, d, e)| \leq \max\{\beta(|x^0|, t), \varepsilon\} \quad \forall t \geq 0$$

whenever  $0 < \varepsilon < |x^0|$ ,  $d(\pi) < \delta(\varepsilon, |x^0|)$ , and for all  $t$ ,  $|e(t)| \leq \eta(\varepsilon, |x^0|)$ , and  $|d(t)| \leq \eta(\varepsilon, |x^0|)$ .

The main result in [18] is as follows.

**Theorem 7** *There is a feedback which stabilizes the system (17) if and only if there is a  $C^1$  clf for the unperturbed system (1).*

It is interesting to note that, as a corollary of Artstein's Theorem, for control-affine systems  $\dot{x} = g_0(x) + \sum u_i g_i(x)$  we may conclude that if there is a discontinuous feedback stabilizer that is robust with respect to small noise, then there is also a regular one, and even one that is smooth on  $\mathbb{R}^n \setminus \{0\}$ . For non control-affine systems, however, there may exist a discontinuous feedback stabilizer that is robust with respect to small noise, yet there is no regular feedback.

Briefly, the sufficiency part of Theorem 7 proceeds by taking a pointwise minimization of the Hamiltonian, for a given  $C^1$  clf, i. e.  $k(x)$  is defined as any  $u$  with  $|u| \leq \sigma(|x|)$  which minimizes  $\nabla V(x) \cdot f(x, u)$ . The necessity part is based on the following technical fact: if the perturbed system can be stabilized, then the differential inclusion

$$\dot{x} \in F(x) := \bigcap_{\varepsilon > 0} \overline{\text{co}} f(x, k(x + \varepsilon B))$$

(where  $B$  denotes the unit ball in  $\mathbb{R}^n$ ) is strongly asymp-

totically stable. One may then apply converse Lyapunov theorems for upper semicontinuous compact convex differential inclusions to deduce the existence of  $V$ .

We now summarize exactly which implications hold, writing “robust” to mean stabilization of the system subject to observation and actuator noise:

$$\begin{array}{ccccc} C^1 V & \iff & \exists \text{ robust } k & & \\ \downarrow & & \downarrow & & \\ C^0 V & \iff & \exists k & \iff & \text{AC} . \end{array}$$

## Future Directions

There are several alternative approaches to feedback stabilization, notably the very appealing approach to discontinuous stabilization through *patchy feedbacks* [1], as well as other related “hybrid” approaches [20]. It is also extremely important to understand the effect of “large” disturbances on the behavior of feedback systems. This study leads one to the very active area of input to state stability (ISS) and related notions (output to state stability as a model of detectability, input to output stability for the study of regulation problems, and so forth), see [27].

## Bibliography

### Primary Literature

1. Ancona F, Bressan A (2003) Flow stability of patchy vector fields and robust feedback stabilization. *SIAM J Control Optim* 41:1455–1476
2. Artstein Z (1983) Stabilization with relaxed controls. *Nonlinear Anal Theory Methods Appl* 7:1163–1173
3. Brockett R (1983) Asymptotic stability and feedback stabilization. In: *Differential geometric control theory*. Birkhauser, Boston, pp 181–191
4. Clarke FH, Ledyaev Y, Sontag E, Subbotin A (1997) Asymptotic controllability implies feedback stabilization. *IEEE Trans Autom Control* 42(10):1394–1407
5. Coron J-M, Rosier L (1994) A relation between continuous time-varying and discontinuous feedback stabilization. *J Math Syst Estim Control* 4:67–84
6. Filippov A (1985) *Differential equations with discontinuous right-hand side*. Nauka, Moscow. English edition: Kluwer, Dordrecht, (1988)
7. Freeman R, Kokotović P (1996) *Robust nonlinear control design*. Birkhäuser, Boston
8. Grüne L, Sontag E, Wirth F (1999) Asymptotic stability equals exponential stability, and ISS equals finite energy gain—if you twist your eyes. *Syst Control Lett* 38(2):127–134
9. Hájek O (1979) Discontinuous differential equations, i & ii. *J Diff Equ* 32(2):149–170, 171–185
10. Hermes H (1967) Discontinuous vector fields and feedback control. In: *Differential equations and dynamical systems*. Proc Internat Sympos, Mayaguez 1965. Academic Press, New York, pp 155–165

11. Isidori A (1995) *Nonlinear control systems: An introduction*. Springer, Berlin
12. Kawski M (1989) Stabilization of nonlinear systems in the plane. *Syst Control Lett* 12:169–175
13. Kawski M (1995) Geometric homogeneity and applications to stabilization. In: Krener A, Mayne D (eds) *Nonlinear control system design symposium (NOLCOS)*, Lake Tahoe, June 1995. Elsevier
14. Krasovskii N, Subbotin A (1988) *Game-theoretical control problems*. Springer, New York
15. Krstić M, Deng H (1998) *Stabilization of uncertain nonlinear systems*. Springer, London
16. Krstić M, Kanellakopoulos I, Kokotović PV (1995) *Nonlinear and adaptive control design*. Wiley, New York
17. Kurzweil J (1956) On the inversion of Ljapunov's second theorem on stability of motion. *Am Math Soc Transl* 2:19–77
18. LedyaeV Y, Sontag E (1999) A Lyapunov characterization of robust stabilization. *Nonlinear Anal* 37(7):813–840
19. Massera JL (1956) Contributions to stability theory. *Ann Math* 64:182–206
20. Prieur C (2006) Robust stabilization of nonlinear control systems by means of hybrid feedbacks. *Rend Sem Mat Univ Pol Torino* 64:25–38
21. Rifford L (2002) Semiconcave control-Lyapunov functions and stabilizing feedbacks. *SIAM J Control Optim* 41:659–681
22. Ryan E (1994) On Brockett's condition for smooth stabilizability and its necessity in a context of nonsmooth feedback. *SIAM J Control Optim* 32:1597–1604
23. Sontag E (1983) A Lyapunov-like characterization of asymptotic controllability. *SIAM J Control Optim* 21(3):462–471
24. Sontag E (1989) A “universal” construction of Artstein's theorem on nonlinear stabilization. *Syst Control Lett* 13(2):117–123
25. Sontag E (1998) *Mathematical control theory*. In: *Deterministic finite-dimensional systems*. Texts in Applied Mathematics, vol 6. Springer, New York
26. Sontag E (1999) Stability and stabilization: Discontinuities and the effect of disturbances. In: *Nonlinear analysis, differential equations and control* (Montreal, 1998). NATO Sci Ser C Math Phys Sci, vol 528. Kluwer, Dordrecht, pp 551–598
27. Sontag E (2006) Input to state stability: Basic concepts and results. In: Nistri P, Stefani G (eds) *Nonlinear and optimal control theory*. Springer, Berlin, pp 163–220
28. Sontag E, Sussmann H (1980) Remarks on continuous feedback. In: *Proc IEEE Conf Decision and Control*, Albuquerque, Dec 1980. pp 916–921

## Books and Reviews

- Clarke FH, LedyaeV YS, Stern RJ, Wolenski P (1998) *Nonsmooth analysis and control theory*. Springer, New York
- Freeman R, Kokotović PV (1996) *Robust nonlinear control design*. Birkhäuser, Boston
- Isidori A (1995) *Nonlinear control systems*, 3rd edn. Springer, London
- Isidori A (1999) *Nonlinear control systems II*. Springer, London
- Khalil HK (1996) *Nonlinear systems*, 2nd edn. Prentice-Hall, Upper Saddle River
- Krstić M, Deng H (1998) *Stabilization of uncertain nonlinear systems*. Springer, London
- Sepulchre R, Janković M, Kokotović PV (1997) *Constructive nonlinear control*. Springer, New York

## Stability Theory of Ordinary Differential Equations

CARMEN CHICONE

Department of Mathematics, University of Missouri-Columbia, Columbia, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Mathematical Formulation of the Stability Concept](#)

[and Basic Results](#)

[Stability in Conservative Systems and the KAM Theorem](#)

[Averaging and the Stability of Perturbed Periodic Orbits](#)

[Structural Stability](#)

[Attractors](#)

[Generalizations and Future Directions](#)

[Bibliography](#)

### Glossary

**Ordinary differential equation** An equation for an unknown vector of functions of a single variable that involves derivatives of the unknown functions. The order of a differential equation is the highest order of the derivatives that appear. The most important class of differential equations are first-order systems of ordinary differential equations that can be written in the form  $\dot{u} = f(u, t)$ , where  $f$  is a given smooth function  $f: U \times J \rightarrow \mathbb{R}^n$ ,  $U$  is an open subset of  $\mathbb{R}^n$ , and  $J$  is an open subset of  $\mathbb{R}$ . The unknown functions are the components of  $u$ , and the vector of their first-order derivatives with respect to the independent variable  $t$  is denoted by  $\dot{u}$ . A solution of this differential equation is a function  $u: K \rightarrow \mathbb{R}^n$ , where  $K$  is an open subset of  $J$  such that  $\frac{du}{dt}(t) = f(u(t), t)$  for all  $t \in K$ .

**Dynamical system** A set and a law of evolution for its elements. The first-order differential equation  $\dot{u} = f(u, t)$ , where  $f: U \times J \rightarrow \mathbb{R}^n$ , is the law of evolution for the set  $U \times J$ ; it defines a continuous (time) dynamical system: Given  $(v, s) \in U \times J$ , the solution  $t \mapsto \psi(t, s, v)$  such that  $\phi(s, s, v) = v$  determines the evolution of the state  $v$ : the state  $v$  at time  $s$  evolves to the state  $\psi(t, s, v)$  at time  $t$ . Similarly, a continuous function  $f: X \rightarrow X$  on a metric space  $X$  defines a discrete dynamical system. The state  $x \in X$  evolves to  $f^k(x)$  (which denotes the value of  $f$  composed with itself  $k$  times and evaluated at  $x$ ) after  $k$  time-steps. The images of  $t \mapsto \phi(t, s, v)$  and  $k \mapsto f^k(x)$  are called the orbits of the corresponding states  $v$  and  $x$ .



**Stability theory** The mathematical analysis of the behavior of the distances between an orbit (or set of orbits) of a dynamical system and all other nearby orbits.

### Definition of the Subject

An orbit or set of orbits of a dynamical system is stable if all solutions starting nearby remain nearby for all future times. This concept is of fundamental importance in applied mathematics: the stable solutions of mathematical models of physical processes correspond to motions that are observed in nature.

### Introduction

Stability theory began with a basic question about the natural world: Is the solar system stable? Will the present configuration of the planets and the sun remain forever; or, might some planets collide, radically change their orbits, or escape from the solar system?

With the advent of Isaac Newton's second law of motion and the law of universal gravitation, the motions of the planets in the solar system were understood to correspond to the solutions of the Newtonian system of ordinary differential equations that modeled the positions and velocities of the planets and the sun according to their mutual gravitational attractions. Short-term approximate predictions (up to a few years in the future) verified this theory; but, due to the complexity of the differential equations of motion, the problem of long-term prediction was not solved; it still occupies a central place in mathematical research.

Does the Newtonian model predict the stability of the solar system?

During the 18th Century, Pierre Simon Laplace [52, 53] asserted a proof of the stability of the solar system. He considered the changes in the semi-major axes and eccentricities of the elliptical motions of the planets around the sun. Using (reasonable) approximations of the Newtonian equations of motion, he showed that for his approximate model these orbital elements do not change over long periods of time due to the disturbances caused by the gravitational attractions of the other bodies in the solar system. If true for the full Newtonian equations of motion, these assertions would imply the stability of the Newtonian solar system. In fact, no proof of the stability of the solar system is known (see [69]). Laplace's results merely provide evidence in favor of stability of the solar system; on the other hand, this work was a primary stimulus for the later development of a general theory of stability.

One of the first rigorous results in stability theory was stated by Joseph-Louis Lagrange [49] and proved by Leje-

une Dirichlet [23,24]; it states that an isolated minimum of the potential energy of a conservative mechanical system is the position of a stable equilibrium point.

Joseph Liouville [56,57,58,59] discussed the problem of the stability of rotating fluid bodies. The further development of this theory was suggested to Aleksandr Mikhailovich Lyapunov as a thesis topic by his advisor Pafnuty Chebyshev. This led to the fundamental and foundational work of Lyapunov on stability theory [63].

Henri Poincaré's introduction of the qualitative theory of differential equations [71] influenced Lyapunov's treatment of stability theory and laid much of the foundation for the modern theory of nonlinear dynamical systems. In addition, Poincaré's work on celestial mechanics [72,73,74,75,76] discusses stability theory.

### Mathematical Formulation of the Stability Concept and Basic Results

Consider the first-order ordinary differential equation

$$\dot{u} = f(u, t), \quad (1)$$

where the dependent variable  $u$  is an  $n$ -dimensional real vector,  $\dot{u}$  denotes the derivative of  $u$  with respect to the independent variable  $t$ , and  $f$  (a mapping, from the cross product of an open subset  $U$  of  $n$ -dimensional space  $\mathbb{R}^n$  and an open subset  $J$  of the real line  $\mathbb{R}$ , into  $\mathbb{R}^n$ ) is at least class  $C^1$ . The existence and uniqueness theorem for differential equations (see, for example, [19]) states that if  $(v, s) \in U \times J$ , then there is a unique solution  $t \mapsto \psi(t, s, v)$  of the differential equation such that  $\psi(s, s, v) = v$ . Moreover, the function  $\psi$  is as smooth as the function  $f$ .

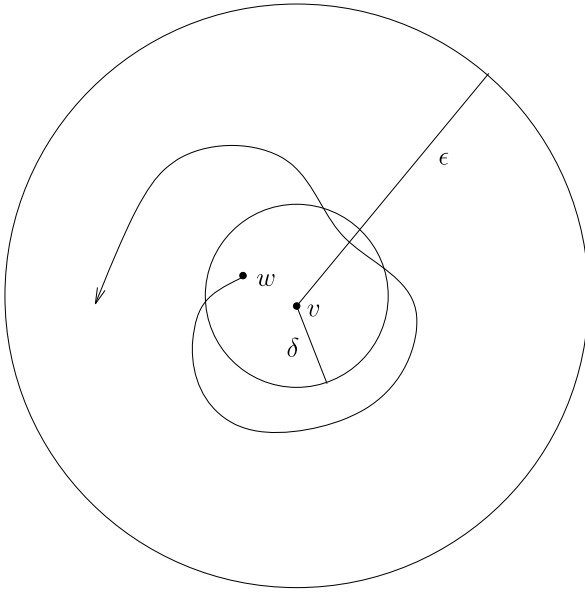
**Definition 1** For  $v \in \mathbb{R}^n$ , the notation  $|v|$  denotes the length of  $v$ .

A solution  $t \mapsto \psi(t, s, v)$  of the differential Eq. (1) is called **stable** if it is defined for all time  $t \geq s$  and for every positive number  $\epsilon$  there is a positive number  $\delta$  such that  $|\phi(t, s, w) - \phi(t, s, v)| < \epsilon$  whenever  $t \geq s$  and  $|w - v| < \delta$  (see Fig. 1).

A **stable solution**  $t \mapsto \psi(t, s, v)$  is called **orbitally asymptotically stable** if there is a choice of  $\delta$  such that the distance between the point  $\psi(t, s, w)$  and the set  $\{\psi(t, s, v) : t \geq s\}$ —called the forward orbit of the solution—converges to zero as  $t \rightarrow \infty$ .

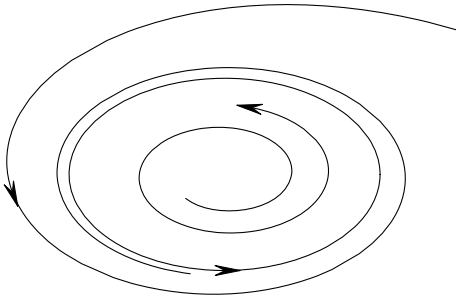
A **stable solution**  $t \mapsto \psi(t, s, v)$  is called **asymptotically stable** if there is a choice of  $\delta$  such that the distance  $|\phi(t, s, w) - \phi(t, s, v)|$  converges to zero as  $t \rightarrow \infty$  whenever  $|w - v| < \delta$  (see Fig. 2).

A solution is called **unstable** if it is not stable.



**Stability Theory of Ordinary Differential Equations, Figure 1**

The figure depicts a stable rest point  $v$ . The orbit starting at  $w$ , an arbitrary point whose distance from  $v$  is less than  $\delta$ , remains within distance  $\epsilon$  for all positive time



**Stability Theory of Ordinary Differential Equations, Figure 2**

The figure depicts an orbitally asymptotically stable elliptical periodic orbit together with two additional orbits that approach the periodic orbit, one from the outside and one from the inside of the region bounded by the periodic orbit

While the definitions just given are widely accepted, the words “asymptotic stability” are often used to mean orbital asymptotic stability. For most situations orbital asymptotic stability is the desired concept. It is clear from the definitions that the concept of asymptotic stability is much stronger than the concept of orbital asymptotic stability for stable solutions whose orbits consist of more than one point. The position of the evolving state along an asymptotically stable orbit is approached by an open set of evolving states. On the other hand, for a solution to be orbitally asymptotically stable only the distances between the ele-

ments of this open set of evolving states and the point set consisting of the forward orbit of the stable solution is required to approach zero. There is no requirement that the open set of evolving states all converge to the evolving state on the stable orbit in unison as time increases without bound. An important concept introduced by Poincaré, the return map, is useful in the study of orbital asymptotic stability; it will be discussed below.

To illustrate the concept of stability, let us consider the (dimensionless) harmonic oscillator

$$\ddot{x} + x = 0. \quad (2)$$

This second-order differential equation is recast as the first-order system

$$\dot{x} = y, \quad \dot{y} = -x \quad (3)$$

by introducing the velocity  $y := \dot{x}$  as a new variable. It is a classical mechanical system with kinetic energy  $\frac{1}{2}\dot{x}^2$  and potential energy  $\frac{1}{2}x^2$ . According to the principle of Lagrange, the equilibrium solution  $(x, y) = (0, 0)$  (which corresponds to an isolated minimum of the potential energy) is stable. This result is easily verified by inspection of the general solution of system (3) that is given by

$$\phi(t, \xi, \eta) = \begin{pmatrix} \cos t & \sin t \\ -\sin t & \cos t \end{pmatrix} \begin{pmatrix} \xi \\ \eta \end{pmatrix}, \quad (4)$$

where  $(\xi, \eta)$  is an arbitrary point in  $\mathbb{R}^2$ . All non-equilibrium solutions are periodic. In fact, the orbit of the solution starting at  $(\xi, \eta)$  is a circle with radius  $\sqrt{\xi^2 + \eta^2}$ . Given an open subset containing the origin, the disk bounded by one of these circles defines an open subset. Every orbit with initial value inside this disk remains inside the disk (and hence the given open set) for all  $t \geq 0$ . On the other hand, the origin is not orbitally asymptotically stable; indeed, the periodic solutions do not approach the origin in positive time.

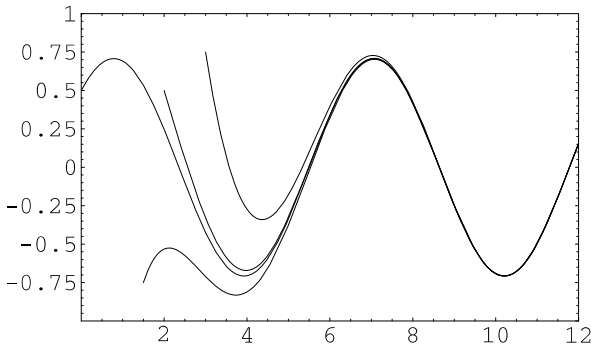
The origin is also an equilibrium of the damped harmonic oscillator

$$\ddot{x} + \epsilon \dot{x} + x = 0, \quad (5)$$

where  $\epsilon > 0$ . Again, inspection of the general solution shows that the origin is asymptotically stable and therefore orbitally asymptotically stable. In effect, all terms in the formula for the general solution contain the factor  $e^{-\epsilon t/2}$ , which converges to zero as  $t \rightarrow \infty$ .

For a nonautonomous example (that is, a differential Eq. (1) such that the partial derivative of  $f$  with respect to  $t$  is not zero), consider the first-order scalar differential equation

$$\dot{x} = -x + \sin t \quad (6)$$



**Stability Theory of Ordinary Differential Equations, Figure 3**

A graph of the asymptotically stable solution (7) of the scalar first-order differential Eq. (6) is depicted together with the graphs of three additional solutions given by  $t \mapsto \phi(t, s, v)$  for  $(s, v)$  equal to  $(1.5, -0.75)$ ,  $(3.0, 0.75)$ , and  $(2.0, 0.5)$

whose solutions are given by

$$\phi(t, s, v) = \left( v - \frac{1}{2}(\sin s + \cos s) \right) e^{s-t} + \frac{1}{2}(\sin t + \cos t).$$

The solution starting at the state  $1/2$  at time zero,

$$\phi\left(t, 0, \frac{1}{2}\right) = \frac{1}{2}(\sin t + \cos t), \quad (7)$$

is asymptotically stable. This result is true due to the presence of the exponential factor in the general solution; it decreases to zero for each fixed  $s$  as  $t \rightarrow \infty$  (cf. Fig. 3).

The origin is an unstable equilibrium for the scalar first-order differential equation  $\dot{x} = x^2$ , whose solutions are given by

$$\phi(t, s, v) = \frac{v}{1 - (t - s)v}.$$

Indeed, for every  $\delta > 0$  (no matter how small) there is a some  $v \in \mathbb{R}$  such that  $|v| < \delta$  and  $v > 0$ . The corresponding solution converges to infinity as  $t \rightarrow (1 + sv)/v$ .

Because explicit solutions of differential equations are rare, the main subject of stability theory is the determination of criteria for stability that do not require knowledge of the general solution of the differential equation.

While the theory of stability is mature with many branches of development, the main results (originally obtained by Poincaré and Lyapunov) follow from two fundamental ideas: linearization and Lyapunov functions.

### The Principle of Linearized Stability

A linear first-order differential equation is a differential equation of the form

$$\dot{u} = A(t)u, \quad (8)$$

where  $A(t)$  denotes an  $n \times n$  matrix-valued function of the independent variable  $t$ .

Most of the analysis of autonomous linear differential equations can be reduced to linear algebra. Indeed, a function of the form  $e^{\lambda t}v$ , where  $\lambda$  is a complex number and  $v \neq 0$  is a complex vector with  $n$  components, is a complex solution of the differential equation

$$\dot{u} = Au, \quad (9)$$

if and only if  $Av = \lambda v$ ; that is,  $\lambda$  is an eigenvalue of the matrix  $A$  and  $v$  is a corresponding eigenvector. The real and imaginary parts of a complex solution are automatically real solutions of the real differential Eq. (9). Also, linear combinations of solutions of linear equations are again solutions. Thus, in case there is a basis of eigenvectors, all solutions of the linear system can be expressed as linear combinations (superpositions) of the special exponential solutions. The general solution, for an arbitrary system matrix  $A$ , can always be obtained from linear combinations of solutions of the form  $p(t)e^{\lambda t}v$ , where  $p(t)$  is a polynomial of degree at most  $n - 1$  and  $Av = \lambda v$ . In fact, there is always a linear change of variables  $w = Bu$ , given by some invertible matrix  $B$ , such that the new system

$$\dot{w} = Jw,$$

where the transformed system matrix  $J := BAB^{-1}$  is in Jordan canonical form (see, for example, [78]). This block diagonal system can be solved by linear combinations of solutions of the form  $p(t)e^{\lambda t}v$ , and the solution of the original system (9) is obtained by the inverse change of variables. Alternatively, it is not difficult to prove that the general solution of the system (9) is given by

$$\phi(t, s, v) = e^{(t-s)A}v, \quad (10)$$

where

$$e^{tA} := I + \sum_{k=1}^{\infty} (tA)^k \quad (11)$$

and  $v$  is an arbitrary point in  $\mathbb{R}^n$ .

Solutions of the form  $t \mapsto p(t)e^{\lambda t}v$  converge to zero as  $t \rightarrow \infty$  whenever the real part of the eigenvalue  $\lambda$  is less than zero. This observation leads to a basic result:

**Theorem 2** *If all eigenvalues of the matrix  $A$  have negative real parts, then the zero solution of the autonomous system (9) is asymptotically stable.*

To examine the stability of a solution  $\psi$  of the (perhaps nonlinear) differential Eq. (1), consider a second solution  $\gamma$ , define the deviation  $\delta = \gamma - \psi$ , and note that

$$\dot{\delta} = f(\gamma, t) - f(\psi, t) = f(\delta + \psi, t) - f(\psi, t)$$

**Definition 3** The linearized equation along the solution  $\psi$  of the differential Eq. (1) is

$$\dot{w} = f_u(\psi(t), t)w, \quad (12)$$

where the subscript denotes the partial derivative with respect to  $u$ .

The linearized equation may be viewed as an approximation of the differential equation for the deviation  $\delta$  because the linearized equation is also obtained by Taylor expanding the function

$$\delta \mapsto f(\delta + \psi, t) - f(\psi, t)$$

to first-order at  $\delta = 0$  and ignoring the remainder term. The principle of linearized stability states that the original solution  $\psi$  is stable whenever the zero solution of the linearized equation is stable.

The principle of linearized stability is not a theorem; rather, it serves as the underlying idea for the theory of stability by linearization. A basic result of this theory is the following theorem (see [63]).

**Theorem 4** *Let*

$$\dot{u} = Au + B(t)u + g(u, t), \quad u(t_0) = u_0, \quad u \in \mathbb{R}^n$$

*be a smooth initial value problem with solution  $t \mapsto \psi(t)$ . If*

- (1) *A is a constant matrix and all its eigenvalues have negative real parts,*
- (2) *B is an  $n \times n$  matrix valued function, continuously dependent on  $t$ , such that the matrix norm  $\|B(t)\|$  converges to zero as  $t \rightarrow \infty$ ,*
- (3) *g is smooth and there are constants  $a > 0$  and  $k > 0$  such that  $|g(v, t)| \leq k|v|^2$  for all  $t \geq 0$  and  $|v| > a$ ,*

*then there are constants  $C > 1$ ,  $\beta > 0$ , and  $\lambda > 0$  such that*

$$|\psi(t)| \leq C|u_0|e^{-\lambda(t-t_0)}, \quad t \geq t_0$$

*whenever  $|u_0| \leq \beta/C$ . In particular, the zero solution is asymptotically stable.*

Theorem 4 is used for the stability analysis of rest points of autonomous first-order differential equations. If  $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$  and  $f(\xi) = 0$ , for some  $\xi \in \mathbb{R}^n$ , then the constant function  $\phi(t, \xi) = \xi$  is a solution of the autonomous differential equation

$$\dot{u} = f(u). \quad (13)$$

In this case,  $\xi$  is called a rest point for the differential equation. By Taylor expansion at the rest point—this time

keeping the remainder term, the differential equation is recast in the form

$$\dot{\gamma} = Df(\xi)\gamma + R(\xi, \gamma),$$

where  $Df$  denotes the derivative of  $f$  and  $\gamma := \phi - \xi$  for an arbitrary solution  $\phi$ . If  $f$  is class  $C^2$ , we have that

$$|R(\xi, v)| \leq k|v|^2.$$

Thus, under the hypothesis that all eigenvalues of the constant system matrix  $Df(\xi)$  have negative real parts, Theorem 4 implies that the rest point  $\xi$  is asymptotically stable. The hypothesis that  $f$  is class  $C^2$  can be relaxed (see, for example, [19]):

**Theorem 5** *If  $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$  is class  $C^1$ ,  $\xi \in \mathbb{R}^n$ ,  $f(\xi) = 0$ , and all eigenvalues of the matrix  $Df(\xi)$  have negative real parts, then the rest point  $\xi$  is asymptotically stable. If the matrix  $Df(\xi)$  has an eigenvalue with positive real part, then the rest point  $\xi$  is unstable.*

The damped harmonic oscillator (5) is equivalent to the first-order system of (linear) differential equations

$$\dot{u} = Au, \quad (14)$$

where the system matrix is

$$A := \begin{pmatrix} 0 & 1 \\ -1 & -\epsilon \end{pmatrix}. \quad (15)$$

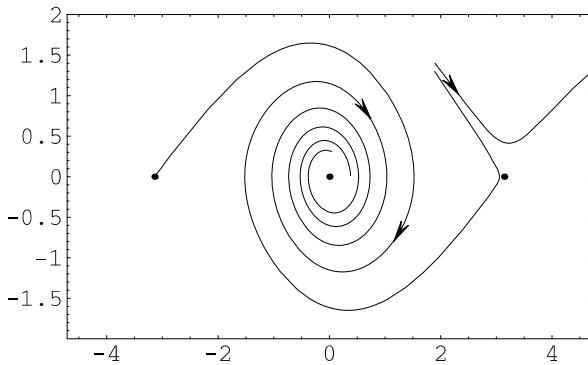
If  $\epsilon > 0$ , the system matrix has eigenvalues  $\frac{1}{2}(-\epsilon \pm \sqrt{\epsilon^2 - 4})$  whose real parts are always negative. Thus, the rest point at the origin is asymptotically stable in agreement with physical intuition.

The power of the linearization method is demonstrated by its applications to systems whose general solutions are not known explicitly; for example, the nonlinear oscillator

$$\ddot{x} + \epsilon x + x + g(x) = 0, \quad (16)$$

where  $g$  is an arbitrary (smooth) function such that  $g(0) = Dg(0) = 0$ . In this case, the origin is a rest point of the equivalent first-order system and the system matrix of the linearized system at the origin is again the matrix (15); therefore, by Theorem 5, the origin is asymptotically stable. This result applies to the rest point  $(\theta, \dot{\theta}) = (0, 0)$  of the damped pendulum

$$\ddot{\theta} + \epsilon \dot{\theta} + \sin \theta = 0. \quad (17)$$



**Stability Theory of Ordinary Differential Equations, Figure 4**

The figure depicts portions of six orbits in the phase portrait of the damped pendulum (17) with  $\epsilon = 0.2$ : the rest points at  $(\theta, \dot{\theta})$  equal to  $(-\pi, 0)$ ,  $(0, 0)$ , and  $(\pi, 0)$ ; and, the orbits starting at  $(-\pi, 0.001)$ ,  $(\pi - 1.25, 1.3)$ , and  $(\pi - 1.243, 1.4)$ . The last two orbits pass close to the (saddle type) unstable rest point at  $(\pi, 0)$

The differential Eq. (17) also has a rest point at  $(\theta, \dot{\theta}) = (\pi, 0)$  where the system matrix of the linearized differential equation is

$$\begin{pmatrix} 0 & 1 \\ 1 & -\epsilon \end{pmatrix}. \quad (18)$$

If  $\epsilon > 0$ , this matrix has an eigenvalue with positive real part; therefore, this rest point—which corresponds to the upward vertical equilibrium of the physical pendulum—is unstable (cf. Fig. 4).

**Definition 6** An  $n \times n$  matrix is called infinitesimally hyperbolic if all of its eigenvalues have non-zero real parts. A rest point of an autonomous system is called hyperbolic if the system matrix of the linearized equation at this rest point is infinitesimally hyperbolic. The rest point is called nondegenerate if the system matrix is invertible.

An important result on the principle of linearized stability for rest points is the Grobman–Hartman theorem (see [30,31,34,35] and [19,22]):

**Theorem 7** If  $v$  is a hyperbolic rest point for the autonomous differential equation  $\dot{u} = f(u)$ , then there is an open set  $V$  containing  $v$  and a homeomorphism  $H$  with domain  $V$  such that the orbits of this differential equation in  $V$  are mapped by  $H$  to orbits of the linearized system  $\dot{w} = Df(v)w$ .

In other words, the qualitative behavior of an autonomous differential equation in a sufficiently small neighborhood of a hyperbolic rest point is the same as the behavior of its linearization at this rest point.

The related problem of the existence of a change of variables (a diffeomorphism) taking a given system to its linearization in a neighborhood of a rest point is more difficult to resolve. Fundamental work in this area is due to Poincaré [74], Carl Siegel [88], Henri Dulac [25], Shlomo Sternberg [90,91] and A.D. Brjuno [16] (see also [9,89]). The non-existence of certain relationships (resonances) among the eigenvalues of the system matrix of the linearization plays an important role in the theory.

**Definition 8** The vector of eigenvalues  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)$  of the  $n \times n$ -matrix  $A$  is called resonant if there is an  $n$ -vector of nonnegative integers  $M = (m_1, m_2, \dots, m_n)$  such that  $\sum_{k=1}^n |m_k| \geq 2$  and at least one of the eigenvalues  $\lambda_k$  is given by

$$\lambda_k = \langle M, \lambda \rangle,$$

where the angle brackets denote the usual inner product. The eigenvalues are nonresonant if no such relationship occurs.

The first theorem of the subject was proved by Poincaré:

**Theorem 9** If the eigenvalues of the linearization of a vector field given by a formal power series are nonresonant, then there is a change of variables in the form of a formal power series that transforms the vector field to its linearization.

In case the linearization is resonant, there is a formal change of variables that removes all of the terms in the formal series that defines the vector field except those that are resonant (that is, the corresponding monomials are given by products of variables whose powers form vectors with the properties of  $M$  in the definition of resonance).

**Definition 10** The vector of eigenvalues  $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_n)$  of the  $n \times n$ -matrix  $A$  belongs to the Poincaré domain if the convex hull of its components in the complex plane does not contain the origin; otherwise, the vector belongs to the Siegel domain. The vector  $\lambda$  is called type  $(C, \nu)$  for  $C > 0$  and  $\nu > 0$ , if the inequality

$$|\lambda_k - \langle M, \lambda \rangle| \geq \frac{C}{(\sum_k |m_k|)^\nu}$$

is satisfied for each component  $\lambda_k$  of  $\lambda$  and all vectors  $M$  as in Definition 8.

**Theorem 11** If the eigenvalues of the linearization of an analytic vector field are nonresonant and in the Poincaré domain, then there is an analytic change of coordinates that transforms the vector field to its linearization. If the eigenvalues of the linearization of an analytic vector field are of



type  $(C, v)$ , then there is an analytic change of coordinates that transforms the vector field to its linearization.

For modern proofs of Poincaré's and Siegel's theorems see [71] or [87]. Sternberg proved Siegel's result for vector fields in the Siegel domain for sufficiently smooth vector fields (see [90,91]).

To successfully apply linearized stability for rest points of autonomous systems, the fundamental problem is to determine the real parts of the eigenvalues of a square matrix; especially, it is important to determine conditions that ensure all eigenvalues have negative real parts. The most important result in this direction is the Routh-Hurwitz criterion (see [40,84]):

**Theorem 12** Suppose that the characteristic polynomial of the real matrix  $A$  is written in the form

$$\lambda^n + a_1\lambda^{n-1} + \cdots + a_{n-1}\lambda + a_n,$$

let  $a_m = 0$  for  $m > n$ , and define the determinants  $\Delta_k$  for  $k = 1, 2, \dots, n$  by

$$\Delta_k := \det \begin{pmatrix} a_1 & 1 & 0 & 0 & 0 & 0 & \cdots & 0 \\ a_3 & a_2 & a_1 & 1 & 0 & 0 & \cdots & 0 \\ a_5 & a_4 & a_3 & a_2 & a_1 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{2k-1} & a_{2k-2} & a_{2k-3} & a_{2k-4} & a_{2k-5} & a_{2k-6} & \cdots & a_k \end{pmatrix}.$$

If  $\Delta_k > 0$  for  $k = 1, 2, \dots, n$ , then all roots of the characteristic polynomial have negative real parts.

The principle of linearized stability is not valid in general. In particular, the stability of the linearized equation at a rest point of an autonomous differential equation does not always determine the stability of the rest point. For example, consider the scalar differential equation  $\dot{u} = -u^m$ , where  $m > 1$  is a positive integer. The origin  $u = 0$  is a rest point for this differential equation and the corresponding linearized system is  $\dot{w} = 0$ . The origin (and every other point on the real line) is a stable rest point of this linear differential equation. On the other hand, by solving the original equation, it is easy to see that the origin is asymptotically stable in case  $m$  is odd and unstable in case  $m$  is even. Thus, the stability of the linearization does not determine the stability of the rest point.

There is no theorem known that can be used to determine the stability of a general nonautonomous first-order linear differential Eq. (8). In particular, there is no obvious relation between the (time-dependent) eigenvalues of a time-dependent matrix and the stability of the zero solution of the corresponding first-order linear differential

equation. Consider, as an example, the  $\pi$ -periodic system  $\dot{u} = A(t)u$ , where

$$A(t) = \begin{pmatrix} -1 + \frac{3}{2}\cos^2 t & 1 - \frac{3}{2}\sin t \cos t \\ -1 - \frac{3}{2}\sin t \cos t & -1 + \frac{3}{2}\sin^2 t \end{pmatrix}, \quad (19)$$

which was constructed by Lawrence Marcus and Hidehiko Yamabe [66]. The real parts of the eigenvalues of  $A(t)$  (given by  $\frac{1}{4}(-1 \pm \sqrt{7}i)$ ) are negative; but,

$$u(t) = e^{t/2} \begin{pmatrix} -\cos t \\ \sin t \end{pmatrix}$$

is a solution that grows without bound.

While a general theory of linearized stability for non-constant solutions of differential equations remains an area of research, much is known in case the non-constant solution is periodic.

Suppose that  $t \mapsto \psi(t)$  is a periodic solution with period  $T$  of the differential Eq. (1) defined on  $\mathbb{R}^n$ . Linearization leads to the differential equation

$$\dot{\gamma} = f_u(\psi(t), t)\gamma, \quad (20)$$

where the matrix (valued function)  $A(t) := f_u(\psi(t), t)$  is periodic with period  $T$ . Thus, the theory of linearized stability for periodic solutions starts with the stability theory for the periodic linear differential Eq. (8). The most basic results in this setting are due to Gaston Floquet [27]. To state them, let us first recall that a linear differential equation admits matrix solutions; that is, matrices whose columns are vector solutions. A fundamental matrix solution is an  $n \times n$ -matrix solution with linearly independent columns; or, equivalently, a matrix solution whose values are all invertible as  $n \times n$ -matrices.

Floquet's main results are stated in the following theorem (see [19] and recall the definition of the matrix exponential in display (11)).

**Theorem 13** If  $\Phi(t)$  is a fundamental matrix solution of the  $T$ -periodic system (8), then

$$\Phi(t + T) = \Phi(t)\Phi^{-1}(0)\Phi(T)$$

for all  $t \in \mathbb{R}$ . In addition, there is a matrix  $B$  (which may be complex) such that

$$e^{TB} = \Phi^{-1}(0)\Phi(T)$$

and a  $T$ -periodic matrix function  $t \mapsto P(t)$  (which may be complex valued) such that  $\Phi(t) = P(t)e^{tB}$  for all  $t \in \mathbb{R}$ . Also, there is a real matrix  $R$  and a real  $2T$ -periodic matrix function  $t \mapsto Q(t)$  such that  $\Phi(t) = Q(t)e^{tR}$  for all  $t \in \mathbb{R}$ .

The stability of the zero solution of the periodic linear differential equation is intimately connected with the eigenvalues of the matrix  $R$  in Floquet's theorem. Indeed, the time-dependent (real) change of variables  $u = Q(t)v$  transforms the  $T$ -periodic differential Eq. (8) to the autonomous linear system

$$\dot{v} = Rv.$$

**Definition 14** The representation  $\Phi(t) = P(t)e^{tB}$  of the fundamental matrix solution  $\Phi$  in Floquet's theorem is called a Floquet normal form. The eigenvalues of the matrix  $e^{tB}$  are called the characteristic (or Floquet) multipliers of the corresponding linear system. A complex number  $\mu$  is called a characteristic (or Floquet) exponent if there is a characteristic multiplier  $\rho$  such that  $e^{\mu T} = \rho$ .

### Theorem 15

- (1) *The characteristic multipliers and the characteristic exponents do not depend on the choice of the fundamental matrix solution of the  $T$ -periodic differential Eq. (8).*
- (2) *If the characteristic multipliers of the periodic system (8) all have modulus less than one; equivalently, if all characteristic exponents have negative real parts, then the zero solution is asymptotically stable.*
- (3) *If the characteristic multipliers of the periodic system (8) all have modulus less than or equal to one; equivalently, if all characteristic exponents have non-positive real parts, and if the algebraic multiplicity equals the geometric multiplicity of each characteristic multiplier with modulus one; equivalently, if the algebraic multiplicity equals the geometric multiplicity of each characteristic exponent with real part zero, then the zero solution is stable.*
- (4) *If at least one characteristic multiplier of the periodic system (8) has modulus greater than one; equivalently, if a characteristic exponent has positive real part, then the zero solution is unstable.*

While Theorem 15 gives a complete description of the stability of the zero solution, no general method is known to determine the characteristic multipliers (cf. [98]). The power of the theorem is an immediate corollary: A finite set of numbers (namely, the characteristic multipliers) determine the stability of the zero solution, at least in the case where none of them have modulus one. The characteristic multipliers can be approximated by numerical integration. And, in special cases, their moduli can be determined by mathematical analysis.

One of the most important applications for Floquet theory and the method of linearization, is the analysis of

Hill's equation

$$\ddot{x} + a(t)x = 0, \quad (21)$$

or, equivalently, the first-order linear system

$$\dot{x} = y, \quad \dot{y} = -a(t)x, \quad (22)$$

where  $x$  is a scalar and  $a$  is a  $T$ -periodic function (see [19, 64]). This equation arose from George Hill's study of lunar motion. It is ubiquitous in stability theory.

The stability of the zero solution of Hill's system (22) can be reduced to the identification of a single number; namely, the magnitude of the trace of the principal fundamental matrix solution  $\Phi(t)$  at  $t = T$  (that is, the fundamental matrix solution  $\Phi(t)$  such that  $\Phi(0) = I$ , where  $I$  is the  $n \times n$  identity matrix).

**Theorem 16** *Suppose that  $\Phi(t)$  is the principal fundamental matrix solution of system (22). If  $|\text{trace } \Phi(T)| < 2$ , then the zero solution is stable. If  $|\text{trace } \Phi(T)| > 2$ , then the zero solution is unstable.*

This result, of course, requires knowledge of the solutions of the system. On the other hand, a beautiful theorem of Lyapunov [55] gives a stability criterion using only properties of the function  $a$ :

**Theorem 17** *If  $a: \mathbb{R} \rightarrow \mathbb{R}$  is a positive  $T$ -periodic function such that*

$$T \int_0^T a(t) dt \leq 4,$$

*then all solutions of the Hill's equation  $\ddot{x} + a(t)x = 0$  are bounded. In particular, the trivial solution is stable.*

The principle of linearized stability for periodic solutions of nonlinear systems motivates a theory similar to the stability theory for rest points. There is, however, one essential difference that will be explained for a  $T$ -periodic solution  $\gamma$  of the autonomous first-order differential equation

$$\dot{u} = f(u). \quad (23)$$

The linearized equation at  $\gamma$  is  $\dot{w} = Df(\gamma(t))w$ . Let  $\phi(t) = f(\gamma(t))$  and note that

$$\dot{\phi}(t) = Df(\gamma(t))f(\gamma(t)) = Df(\gamma(t))\phi(t);$$

that is,  $t \mapsto f(\gamma(t))$  is a solution of the linearized equation. This solution is given by

$$f(\gamma(t)) = \Phi(t)f(\gamma(0)),$$

where  $\Phi(t)$  is the principal fundamental matrix solution of the linearized equation. It follows that  $\Phi(T)f(\gamma(0)) =$

$f(\gamma(0))$ ; therefore, the number 1 is a characteristic multiplier. It corresponds to motion in the direction of the periodic solution and does not contribute to the stability or instability of the periodic solution. This characteristic multiplier must be treated separately in the stability theory.

**Theorem 18** *If the number 1 occurs with algebraic multiplicity one in the set of characteristic multipliers of the linearized differential equation at a periodic solution and all other characteristic multipliers have modulus less than one, then the periodic solution is orbitally asymptotically stable. If at least one characteristic multiplier has modulus larger than one, then the periodic solution is unstable.*

Poincaré introduced a geometric approach to stability for periodic solutions. His idea is simple and useful: Consider a periodic solution  $\gamma$  and some point, say  $\gamma(0)$ , on the orbit of this solution in  $\mathbb{R}^n$ . Construct a hypersurface  $S$  in  $\mathbb{R}^n$  that contains  $\gamma(0)$  and is transverse to the orbit of  $\gamma$  at this point. By the implicit function theorem, there is an open neighborhood  $\Sigma$  of  $\gamma(0)$  in  $S$  such that the solution of the differential equation starting at every point in  $\Sigma$  returns to  $S$ . This defines a local diffeomorphism  $\mathcal{P}: \Sigma \rightarrow S$ , called the Poincaré (or return) map, which assigns  $p$  to its point of first return on  $S$  (cf. Fig. 5).

The eigenvalues of the derivative of  $\mathcal{P}$  at  $\gamma(0)$  are exactly the characteristic multipliers associated with  $\gamma$ . Thus, if all eigenvalues of this derivative have modulus less than one, then  $\gamma$  is orbitally asymptotically stable. The derivative of the Poincaré map can be computed using the linearized equations of the corresponding differential equation along its the periodic orbit. While explicit formulas for the desired derivative are only available for special cases, the Poincaré map provides an indispensable tool for

organizing the study of stability for periodic orbits in general.

It should be clear that the discrete dynamical system defined by  $\mathcal{P}$  codes all the information about the qualitative behavior of the solutions of the differential equation that start near the orbit of  $\gamma$ . The existence of the Poincaré map is thus the main connection between discrete and continuous dynamical systems.

Often, but not always, the dynamical properties of diffeomorphisms (or maps) are simpler to analyze than the corresponding properties of differential equations. Thus, theorems are often proved first for maps and later for differential equations. For instance, the Grobman–Hartman theorem was first proved for diffeomorphisms. This result can be applied to the Poincaré map of a periodic orbit to prove the existence of a homeomorphism that maps nearby orbits to the orbits of a corresponding linearization.

For the special case of periodic solutions of autonomous first-order differential equations on the plane, there is exactly one important Floquet multiplier; it is identified in the following fundamental result.

**Theorem 19** *If  $\gamma$  is a periodic solution with period  $T$  of the differential Eq. (23) defined on  $\mathbb{R}^2$ , then the number*

$$\mu = \int_0^T \operatorname{div} f(\gamma(t)) dt$$

(where  $\operatorname{div}$  denotes the divergence) is (up to a positive scalar multiple) a characteristic exponent of  $\gamma$ . If  $\mu$  is negative, then  $\gamma$  is orbitally asymptotically stable. If  $\mu$  is positive, then  $\gamma$  is unstable.

The system of differential equations

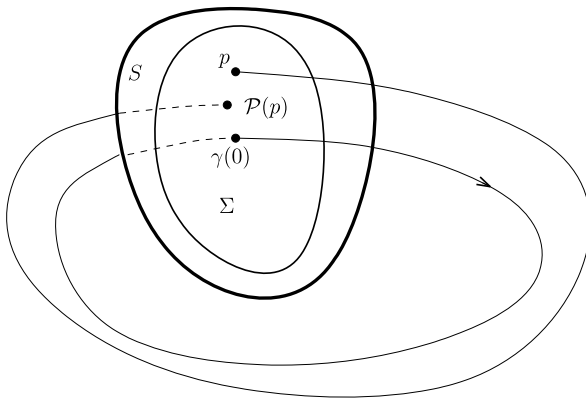
$$\dot{x} = x - y - (x^2 + y^2)x, \quad \dot{y} = x + y - (x^2 + y^2)y$$

has the periodic solution  $t \rightarrow (\cos t, \sin t)$  (depicted with a non-unit aspect ratio in Fig. 2). Also, the divergence of the vector field is  $2 - 4(x^2 + y^2)$ ; it has value -2 along the corresponding periodic orbit  $\Gamma = \{(x, y): x^2 + y^2 = 1\}$ . By Theorem 19,  $\Gamma$  is orbitally asymptotically stable. This is an example of a limit cycle; that is, an isolated periodic orbit. By changing to polar coordinates  $(r, \theta)$ , the transformed system

$$\dot{r} = r(1 - r^2), \quad \dot{\theta} = 1$$

decouples, and its flow is given by

$$\phi_t(r, \theta) = \left( \left( \frac{r^2 e^{2t}}{1 - r^2 + r^2 e^{2t}} \right)^{\frac{1}{2}}, \theta + t \right). \quad (24)$$



**Stability Theory of Ordinary Differential Equations, Figure 5**  
A schematic diagram of a Poincaré map  $\mathcal{P}$  defined on a Poincaré section  $\Sigma$  near a periodic orbit containing the point  $\gamma(0)$

The positive  $x$ -axis is a Poincaré section with corresponding Poincaré map  $P$  given by

$$P(x) = \left( \frac{x^2 e^{4\pi}}{1 - x^2 + x^2 e^{4\pi}} \right)^{\frac{1}{2}}.$$

The periodic orbit  $\Gamma$  intersects the Poincaré section at  $x = 1$  where we have  $P(1) = 1$  and  $P'(1) = e^{-4\pi} < 1$ .

A periodic solution of an autonomous system cannot be asymptotically stable. This fact depends on a special property of the solutions of autonomous systems: if the differential Eq. (1) is autonomous, then the function  $\psi$  giving the solutions,  $t \mapsto \psi(t, s, v)$ , is a function of  $t - s$ . Hence, by replacing  $t - s$  by  $t$ , each solution can be expressed in the form  $t \mapsto \phi(t, v)$  where  $\phi(0, v) = v$  and  $\phi(t, \phi(s, v)) = \phi(t + s, v)$  whenever both sides are defined. The family of functions  $v \mapsto \phi(t, v)$ , parametrized by  $t$  is called the flow of the differential Eq. (1).

Suppose that  $t \mapsto \phi(t, v)$  is an asymptotically stable periodic solution with period  $T$ . Let  $\delta > 0$  be as in Definition 1. If  $|w - v| < \delta$ , then  $\lim_{t \rightarrow \infty} |\phi(t, w) - \phi(t, v)| = 0$ . For all sufficiently small  $s$ , we have that  $|\phi(s, w) - v| < \delta$ . Hence,  $\lim_{t \rightarrow \infty} |\phi(t, \phi(s, w)) - \phi(t, v)| = 0$ . By the triangle law,

$$\begin{aligned} & |\phi(t + s, v) - \phi(t, v)| \\ & \leq |\phi(t + s, v) - \phi(t + s, w)| + |\phi(t + s, w) - \phi(t, v)| \\ & \leq |\phi(s, \phi(t, v)) - \phi(s, \phi(t, w))| + |\phi(t + s, w) - \phi(t, v)|. \end{aligned}$$

Since  $\xi \mapsto \phi(s, \xi)$  is smooth and a periodic orbit is compact, the law of the mean implies that there is a constant  $C > 0$  such that

$$\begin{aligned} & |\phi(t + s, v) - \phi(t, v)| \\ & \leq C|\phi(t, v) - \phi(t, w)| + |\phi(t + s, w) - \phi(t, v)|. \end{aligned}$$

Both summands on the right-hand side of the last inequality converge to zero as  $t \rightarrow \infty$ . Hence, we have that

$$\lim_{k \rightarrow \infty} |\phi(kT + s, v) - \phi(kT, v)| = 0,$$

where  $k$  denotes an integer variable. It follows that  $\phi(s, v) = v$  for all  $s$  in some open set. The point  $v$  must be a rest point on a periodic orbit, in contradiction to the uniqueness of solutions.

While a periodic solution of an autonomous system cannot be asymptotically stable, we can ask that an orbitally asymptotically stable periodic orbit satisfy a stronger type of stability.

**Definition 20** Let  $\phi$  denote the flow of the autonomous differential Eq. (23) and suppose that  $\Gamma$  is a periodic orbit corresponding to the solution  $t \mapsto \phi(t, p)$ . We say that

$q \in \mathbb{R}^n$  has asymptotic phase  $p$  with respect to  $\Gamma$  if the solution  $t \mapsto \phi(t, q)$  is such that

$$\lim_{t \rightarrow \infty} |\phi(t, q) - \phi(t, p)| = 0.$$

We say that  $\Gamma$  is isochronous if there exists an open neighborhood of its orbit such that every point in this neighborhood has asymptotic phase with respect to the period solution.

**Theorem 21** *If all eigenvalues of the derivative of a Poincaré map at the periodic solution  $t \mapsto \phi(t, p)$  of (23) lie strictly inside the unit circle in the complex plane (equivalently, the corresponding periodic orbit is hyperbolic), then every point  $q$  in some open neighborhood of this periodic orbit has asymptotic phase.*

A theory of asymptotic phase that includes orbitally stable periodic orbits such that some eigenvalues of the derivative of an associated Poincaré map have modulus one was recently completed (see [21,26]).

## Lyapunov Functions

In case the linearized differential equation along a solution is stable but not orbitally asymptotically stable, the principle of linearized stability usually cannot be justified. One of the great contributions of Lyapunov is the introduction of a method that can be used to determine the stability of such solutions.

The fundamental ideas of Lyapunov are most clear for the stability analysis of rest points of autonomous systems (see [63]); but, the theory goes far beyond this case (see [51]).

**Definition 22** Let  $u_0$  be a rest point of the autonomous differential Eq. (23). A continuous function  $L: U \rightarrow \mathbb{R}$ , where  $U \subseteq \mathbb{R}^n$  is an open set with  $u_0 \in U$ , is called a Lyapunov function at  $u_0$  if

- (1)  $L(u_0) = 0$ ,
- (2)  $L(x) > 0$  for  $x \in U \setminus \{u_0\}$ ,
- (3) the function  $L$  is continuously differentiable on the set  $U \setminus \{u_0\}$ , and, on this set,  $\dot{L}(x) := \text{grad } L(x) \cdot f(x) \leq 0$ .

The function  $L$  is called a strict Lyapunov function if, in addition,

- (4)  $\dot{L}(x) < 0$  for  $x \in U \setminus \{u_0\}$ .

Property (3) states that  $L$  does not increase along solutions.

**Theorem 23** *If there is a Lyapunov function defined on an open neighborhood of a rest point of an autonomous first-order differential equation, then the rest point is stable. If, in*

addition, the Lyapunov function is a strict Lyapunov function, then the rest point is asymptotically stable.

Lyapunov's theorem was motivated by Lagrange's principle: a rest point of a mechanical system corresponding to an isolated minimum of the potential energy is stable. To obtain this result, recall that the equation of motion of a (conservative) mechanical system for the motion of a particle is

$$m\ddot{x} = -\text{grad } G(x)$$

where  $m$  is the mass of the particle and  $x \in \mathbb{R}^n$  is its position. The kinetic energy of the particle is defined to be  $K = \frac{m}{2} \langle \dot{x}, \dot{x} \rangle$  (where the angle brackets denote the usual inner product) and the potential energy is  $G$ , a quantity defined up to an additive constant.

Consider the corresponding equivalent first-order system

$$\dot{x} = y, \quad \dot{y} = -\frac{1}{m} \text{grad } G(x),$$

and suppose that  $x_0$  be an isolated minimum of  $G$ . The state  $(x, y) = (x_0, 0)$  is a rest point of the mechanical system. By inspection, the total energy

$$L(x, y) := \frac{m}{2} \langle y, y \rangle + G(x) - G(x_0)$$

(with an appropriate translation of the potential energy) is a Lyapunov function at the rest point. Thus, the rest point is stable.

The proof of the first part of Lyapunov's theorem is simple: Choose an open ball  $B$  with center at the rest point and radius sufficiently small so that its closure is in the domain of the Lyapunov function  $L$ . The continuous positive function  $L$  on the closed and bounded spherical boundary of  $B$  has a non-zero minimum value  $b$ . Because the continuous function  $L$  vanishes at the rest point, there is a second ball  $A$  that is concentric with  $B$ , has strictly smaller radius, and is such that the maximum value of  $L$  on all of  $A$  is smaller than  $b$ . A solution of the differential equation starting in  $A$  cannot reach the boundary of  $B$  because  $L$  does not increase along solutions. Thus, the rest point is stable.

Lyapunov's method extends to nonautonomous systems with an appropriate modification of the notion of a Lyapunov function. Suppose that the first-order differential Eq. (1) has a rest point in the sense that for some  $T \in \mathbb{R}$ ,  $f(u_0, t) = 0$  for all  $t \geq T$ . A Lyapunov function is a class  $C^1$  function defined in a neighborhood of  $u_0$  for all  $t \geq T$  such that  $L(u_0, t) = 0$  for  $t \geq T$ , there is a continuous nonnegative function  $M$  defined on the same neighborhood of  $u_0$  such that  $L(u, t) \geq M(u)$  for all  $t \geq T$

and  $\dot{L} = L_t + L_u \cdot \dot{u} \leq 0$  for all  $t \geq T$ . If such a Lyapunov function exists, then the rest point is stable. If, in addition, there is a positive function  $N$  with the same domain as  $M$  such that  $\dot{L}(u, t) \leq -N(u)$  for all  $t \geq T$ , then the rest point is asymptotically stable.

Lyapunov's direct method can be used to prove the principle of linearized stability for rest points of autonomous systems. To see how this might be done, consider the differential equation

$$\dot{u} = Au + g(u), \quad u \in \mathbb{R}^n,$$

where  $A$  is a real  $n \times n$  matrix and  $g: \mathbb{R}^n \rightarrow \mathbb{R}^n$  is a smooth function. Suppose that every eigenvalue of  $A$  has negative real part, and that for some  $a > 0$ , there is a constant  $k > 0$  such that, using the usual norm on  $\mathbb{R}^n$ ,

$$|g(x)| \leq k|x|^2$$

whenever  $|x| < a$ . Let  $\langle \cdot, \cdot \rangle$  denote the usual inner product on  $\mathbb{R}^n$ , and let  $A^*$  denote the transpose of the real matrix  $A$ . Suppose that there is a real symmetric positive definite  $n \times n$  matrix that also satisfies Lyapunov's equation

$$A^*B + BA = -I$$

and define  $L: \mathbb{R}^n \rightarrow \mathbb{R}$  by

$$L(x) = \langle x, Bx \rangle.$$

Using Schwarz's inequality, it can be proved that the restriction of  $L$  to a sufficiently small neighborhood of the origin is a strict Lyapunov function. The proof is completed by showing that there is a symmetric positive-definite solution of Lyapunov's equation. In fact,

$$B := \int_0^\infty e^{tA^*} e^{tA} dt$$

is a symmetric positive definite  $n \times n$  matrix that satisfies Lyapunov's equation. Alternatively, it is possible to prove the existence of a solution to Lyapunov's equation using purely algebraic methods (see, for example, [97]).

The instability of solutions can also be detected by Lyapunov's methods. A simple result in this direction is the content of the following theorem.

**Theorem 24** Suppose that  $L$  is a smooth function defined on an open neighborhood  $U$  of the rest point  $u_0$  of the autonomous differential Eq. (23) such that  $L(u_0) = 0$  and  $\dot{V}(u) > 0$  on  $U \setminus \{u_0\}$ . If  $L$  has a positive value somewhere in each open set containing  $u_0$ , then  $u_0$  is unstable.

One indication of the subtlety of the determination of stability is the unsolvability of the center-focus problem [41].



Consider a planar system of first-order differential equations in the form

$$\dot{x} = y + P(x, y), \quad \dot{y} = -x + Q(x, y),$$

where  $P$  and  $Q$  are analytic at the origin with leading-order terms at least quadratic in  $x$  and  $y$ . Is the origin a focus (that is, asymptotically stable or asymptotically unstable) or a center (that is, all orbits in some neighborhood of the origin are periodic)? There is no algorithm that can be used to solve this problem for all such systems in a finite number of steps. On the other hand, the center-focus problem is solved, by using Lyapunov's stability theorem, if a certain sequence of numbers called Lyapunov quantities (which can be computed iteratively and algebraically) has a non-zero element. One problem is that the algorithm for computing the Lyapunov quantities may not terminate in a finite number of steps (see, for example, [19,89]).

The center-focus problem is solved for some special cases. The most important theorem in this area of research is due to Nikolai Bautin (see [14] and [99,100] for a modern proof); he found (among other results) a complete solution of the center-focus problem for quadratic systems (that is, where  $P$  and  $Q$  are homogeneous quadratic polynomials). While the center-focus problem for quadratics had been solved by Dulac [25] and others [28,43,44,45] before his paper appeared; Bautin introduced a method, which involves the use of polynomial ideals, that has had far reaching consequences.

### Stability in Conservative Systems and the KAM Theorem

Stability theory for conservative systems leads to many delicate problems, some of which—for example Lagrange's problem—can be resolved by the method of Lyapunov functions. While the total energy, the total angular momentum, or other first integrals all have the property that their derivatives vanish along trajectories, these integrals do not always serve as Lyapunov functions because they often fail to be positive definite in punctured neighborhoods of rest points or periodic orbits. Thus, other methods are required.

As in the quest to determine the stability of the solar system, a basic problem in mechanics is to determine the stability of periodic motions (or rest points) of conservative differential equations with respect to small perturbations. While no general solution is known, great progress has been made culminating in the Kolmogorov–Arnold–Moser (KAM) theorem (see [5,46,47,68,87,89,92]).

A mechanical system with  $N$  degrees-of-freedom (that is, the positions are determined by  $N$ -coordinates) is called

completely integrable if it has  $N$  independent first integrals whose Poisson brackets are in involution (see p. 271 in [8]). In this case, the system can be transformed to a first-order system of differential equations in the simple form

$$\dot{I} = 0, \quad \dot{\theta} = \Omega(I), \quad (25)$$

where  $I$  is an  $N$ -dimensional variable (of actions),  $\theta$  is an  $N$ -dimensional variable (of angles), and  $\Omega$  is a smooth function. The new coordinates are called action-angle variables (see, for generalizations, [39,42]). This system is in Hamiltonian form ( $\dot{I} = \partial H / \partial \theta$ ,  $\dot{\theta} = -\partial H / \partial I$ ) with Hamiltonian  $H(I, \theta) := -\omega(I)$ , where  $\omega$  is an anti-derivative of  $\Omega$ ; and, of course, the Hamiltonian is constant along solutions of the system.

The corresponding perturbed system is taken to have the form

$$\dot{I} = \epsilon \frac{\partial F}{\partial \theta}(I, \theta), \quad \dot{\theta} = \Omega(I) - \epsilon \frac{\partial F}{\partial I}(I, \theta) \quad (26)$$

(or, more generally, a similar form with additional terms that are higher-order in  $\epsilon$ ), where  $F$  is a smooth function that is  $2\pi$ -periodic in  $\theta$ . This system is in Hamiltonian form with Hamiltonian

$$\mathcal{H}(I, \theta, \epsilon) := -\omega(I) + \epsilon F(I, \theta).$$

The unperturbed integrable system can be solved explicitly as

$$I(t) = I_0, \quad \theta(t) = \Omega(I_0)t + \theta_0;$$

hence, by inspection, all orbits are periodic or quasi-periodic (when the angular variables are defined modulo  $2\pi$ ) according to whether or not the vector of frequencies  $\Omega(I_0)$  satisfies a resonance relation, that is,  $\langle M, \Omega(I_0) \rangle = 0$  for some  $N$ -vector  $M$  with integer components.

Every orbit of the unperturbed system with one degree-of-freedom is periodic; but, for example for two degrees-of-freedom, orbits are periodic with period  $T$  only if there are integers  $m_1$  and  $m_2$  such that

$$T\Omega_1(I_0) = 2\pi m_1, \quad T\Omega_2(I_0) = 2\pi m_2$$

$$\text{or } m_1\Omega_2(I_0) - m_2\Omega_1(I_0) = 0.$$

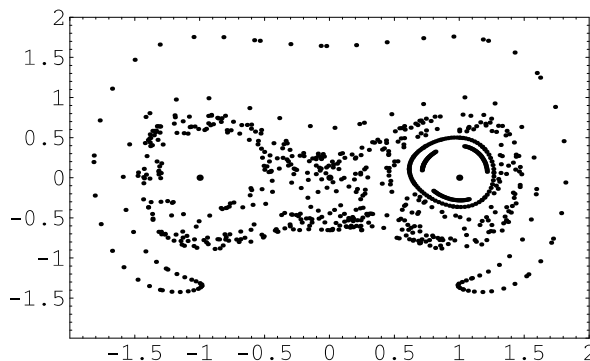
Geometrically, a choice of actions fixes an energy level and the angles correspond to an  $N$ -dimensional invariant torus. The flow on this torus is either periodic or quasi-periodic according to the existence of a corresponding resonance relation. For this reason, the tori with periodic flows

are called resonant; the others are called nonresonant. The KAM theorem states that most of the nonresonant tori in the unperturbed system (25) survive after small perturbations as invariant tori of the perturbed system (26).

**Theorem 25** *If (in the Hamiltonian system (26))  $\mathcal{H}$  is sufficiently smooth,  $D\Omega(I)$  is invertible, and  $\epsilon$  is sufficiently small, then almost all (in the sense of Lebesgue measure) nonresonant unperturbed invariant tori persist.*

There are similar theorems for time-dependent perturbations of integrable systems and for area-preserving maps (see [8]).

For a mechanical system with  $N$  degrees-of-freedom, the perturbed invariant tori are  $N$ -dimensional manifolds embedded in the  $2N$ -dimensional phase space. After restriction to a regular energy surface (that is, a level set of  $\mathcal{H}$  corresponding to one of its regular values), these  $N$ -dimensional tori are embedded in a  $(2N - 1)$ -dimensional manifold. For  $N \leq 2$ , the invariant tori that exist by the KAM theorem separate the energy hypersurface into two components, one inside and one outside of each invariant torus. A motion starting in a bounded region whose boundary is one of these invariant tori cannot escape to the corresponding unbounded component. Since the invariant tori are dense, such motions are stable (cf. Fig. 6). More generally, each perturbed invariant torus is stable as an invariant set. This result can be used, for example, to prove the stability of certain periodic orbits in the restricted three-body problem (see [54]). On the other hand,



**Stability Theory of Ordinary Differential Equations, Figure 6**

The figure depicts portions of seven orbits of the (stroboscopic) Poincaré map  $\xi \in \mathbb{R}^2 \mapsto \phi(2\pi, \xi)$  for the forced oscillator  $\dot{x} = y$ ,  $\dot{y} = x - x^3 + 0.05 \sin t$ : three fixed points at  $(x, y)$  equal to  $(-1, 0)$ ,  $(0, 0)$ , and  $(1, 0)$ ; an orbit starting at  $(x, y) = (1.0, 0.4)$  at time  $t = 0$ , which is close to a  $(3 : 1)$  resonance; an orbit starting at  $(1.0, 0.5)$ , which is on an invariant torus that surrounds the resonant orbit; an orbit starting at  $(1.0, 0.6)$ , which appears to be “chaotic”; and an orbit starting at  $(1, 1)$ , which is on a large invariant torus. The  $(3 : 1)$  resonant orbit is stable

for  $N > 2$ , the invariant tori no longer separate energy surfaces into bounded and unbounded components. Thus, it is possible for orbits to migrate outside nearby tori in a conjectured process called Arnold diffusion (see [6]). The validity of Arnold diffusion for the general Hamiltonian system (26) is an area of current research (see, for a review, [60]).

### Averaging and the Stability of Perturbed Periodic Orbits

As we have seen, systems of the form

$$\dot{I} = \epsilon P(I, \theta), \quad \dot{\theta} = \Omega(I) + \epsilon Q(I, \theta), \quad (27)$$

where  $P$  and  $Q$  are  $2\pi$ -periodic, often arise in mechanics. Here, the properties of system (27) are considered for general perturbations, which are not necessarily conservative.

Note that, in case  $\epsilon$  is small, the vector of actions (components of  $I$ ) in the differential Eq. (27) evolves slowly relative to the evolution of the vector of angles  $\theta$ . The averaging principle states that the evolution of the actions for such a system is well-approximated by the corresponding averaged equation given by

$$\dot{I} = \epsilon \bar{P}(I), \quad (28)$$

where

$$\bar{P}(I) := \frac{1}{(2\pi)^N} \int_{\mathbb{T}^N} P(I, \theta) d\theta$$

and  $\mathbb{T}^N$  denotes the  $N$ -dimensional torus of angles.

Exactly what is meant by “well-approximated” is clarified by the averaging theorem, which requires a severe restriction: the vector of angles is one-dimensional (cf. [61, 86]).

**Theorem 26** *If  $\theta$  is a scalar variable and  $\Omega$  is bounded away from zero, then there is a near-identity change of variables of the form  $\mathcal{I} = I + \epsilon k(I, \theta)$  that is  $2\pi$ -periodic in  $\theta$  which transforms system (27) into the form*

$$\dot{\mathcal{I}} = \epsilon \bar{P}(\mathcal{I}) + \epsilon^2 P_1(\mathcal{I}, \theta, \epsilon), \quad \dot{\theta} = \Omega(\mathcal{I}) + \epsilon Q_1(\mathcal{I}, \theta, \epsilon).$$

where  $P_1$  and  $P_2$  are  $2\pi$ -periodic in  $\theta$ . Moreover, if the evolution of  $I$  starting at  $I_0$  is given by  $I(t)$  and the evolution (according to the differential Eq. (28)) of the averaged action  $\mathcal{I}$  starting at  $I_0$  is given by  $\mathcal{I}(t)$ , then there are constants  $C > 0$  and  $T(I_0) > 0$  such that

$$|\mathcal{I}(t) - I(t)| \leq C\epsilon$$

on the time interval  $0 \leq \epsilon t < T(I_0)$ .

The existence and stability of perturbed period solutions is the content of the following theorem.

**Theorem 27** Consider the system

$$\dot{I} = \epsilon F(I, \theta) + \epsilon^2 F_2(I, \theta, \epsilon), \quad \dot{\theta} = \Omega(I) + \epsilon G(I, \theta, \epsilon), \quad (29)$$

where  $I \in \mathbb{R}^M$ ,  $\theta$  is a scalar,  $F$ ,  $F_2$ , and  $G$  are  $2\pi$ -periodic functions of  $\theta$ , and there is some number  $c$  such that  $\Omega(I) \geq c > 0$ . If the averaged system has a nondegenerate rest point (see Definition 6) and  $\epsilon$  is sufficiently small, then system (29) has a periodic orbit. If in addition  $\epsilon > 0$  and the rest point is hyperbolic, then the periodic orbit has the same stability type as the hyperbolic rest point; that is, the dimensions of the corresponding stable and unstable manifolds are the same (see Definition 29).

For a typical application of Theorem 27, consider the system of differential equations

$$\begin{aligned} \dot{I} &= \epsilon(a + b \sin k\theta) \sin \psi + \epsilon^2 P(I, \psi, \theta), \\ \dot{\psi} &= \epsilon c I + \epsilon^2 Q(I, \psi, \theta), \\ \dot{\theta} &= 1, \end{aligned} \quad (30)$$

where  $a$ ,  $b$ , and  $c$  are non-zero constants,  $k$  is a positive integer, and both  $P$  and  $Q$  are  $2\pi$ -periodic in  $\theta$ . The averaged system of differential equations

$$\dot{I} = \epsilon a \sin \psi, \quad \dot{\psi} = \epsilon c I$$

has hyperbolic rest points at  $(I, \psi) = (0, 0)$  and  $(I, \psi) = (0, \pi)$ . According to Theorem 27, if  $\epsilon$  is sufficiently small, the system of differential Eq. (30) has corresponding hyperbolic periodic solutions. In case  $ac > 0$ , the solution of the perturbed system corresponding to  $\psi = 0$  is stable, and the solution corresponding to  $\psi = \pi$  is unstable (in fact, it is a hyperbolic saddle). If  $ac < 0$ , the stability types are switched.

### Structural Stability

An important aspect of the global theory of dynamical systems is the stability of the orbit structure as a whole. The motivation for the corresponding theory comes from applied mathematics. Mathematical models always contain simplifying assumptions. Dominant features are modeled; supposed small disturbing forces are ignored. Thus, it is natural to ask if the qualitative structure of the set of solutions—the phase portrait—of a model would remain the same if small perturbations were included in the model. The corresponding mathematical theory is called structural stability.

The correct mathematical formulation of the definition of structural stability requires the introduction of a topology on the space of dynamical systems under consideration. The natural setting for the theory is the space of class  $C^1$  vector fields on a smooth compact manifold.

There is a one-to-one correspondence between vector fields and differential equations: the right-hand side of a first-order autonomous differential equation defines a vector field on  $\mathbb{R}^n$ . The local representatives, with respect to coordinate charts, of a vector field on an  $n$ -dimensional manifold  $M$  are vector fields on  $\mathbb{R}^n$ . The set of all smooth vector fields  $\mathcal{X}(M)$  on  $M$  has the structure of a Banach space after the introduction of a Riemannian metric. Thus, two vector fields are close if the distance between them in this Banach space is small.

**Definition 28** A vector field  $X$  on a smooth manifold  $M$  is called structurally stable, if there is some neighborhood  $U$  of  $X$  in  $\mathcal{X}(M)$  such that for each  $Y \in U$  there is a homeomorphism of  $M$  that maps orbits of  $Y$  to orbits of  $X$  and preserves the orientation of orbits according to the direction of time. Such a homeomorphism is called a topological equivalence.

In other words, all perturbations of a structurally stable vector field have the same qualitative orbit structures.

At first impression, it might seem that the homeomorphism in the definition of structural stability should be a diffeomorphism. But, this requirement is too strong. To see why, consider a hyperbolic rest point  $u_0$  of the first-order autonomous differential Eq. (23) in  $\mathbb{R}^n$ . Let  $\dot{u} = g(u)$  be a differential equation on  $\mathbb{R}^n$ , and consider the perturbation of the differential Eq. (23) given by  $\dot{u} = f(u) + \epsilon g(u)$ . We have that  $f(u_0) = 0$ . Thus,  $F(u, \epsilon) := f(u) + \epsilon g(u)$  is such that  $F(u_0, 0) = 0$  and  $DF(u_0, 0) = Df(u_0)$ . Because of the hyperbolicity, the linear transformation  $Df(u_0)$  is invertible. By an application of the implicit function theorem, there is a smooth function  $\beta: J \rightarrow \mathbb{R}^n$ , where  $J$  is an open interval in  $\mathbb{R}$  containing the origin, such that  $F(\beta(\epsilon), \epsilon) = 0$  for all  $\epsilon$  in  $J$ . In other words, every small perturbation of (the vector field)  $f$  has a rest point near  $u_0$ . Moreover, for sufficiently small  $\epsilon$  (and in view of the continuity of eigenvalues with respect to the components of the corresponding matrix), the perturbed rest point  $\beta(\epsilon)$  is hyperbolic with the same number of eigenvalues (counting multiplicities) with positive and negative real parts as the eigenvalues of  $Df(u_0)$ . But, of course, the perturbed eigenvalues are in general not the same as the eigenvalues of  $Df(u_0)$ . The hyperbolic rest point  $u_0$  is structurally stable in the sense that the local phase portraits at the corresponding rest points of sufficiently small perturbations of  $f$  are qual-

itatively the same as the phase portrait of the differential Eq. (23). This fact can be proved using the Grobman–Hartman theorem together with an argument to show that the phase portraits of hyperbolic linear first-order systems are topologically equivalent if their system matrices have the same numbers of eigenvalues (counting multiplicities) with positive and negative real parts. Suppose there were a diffeomorphism  $h$  taking orbits to orbits. It would correspond to a change of variables  $v = h(u)$  (a smooth conjugacy) taking the differential Eq. (23) to the differential equation  $\dot{v} = F(v)$ , where the parameter  $\epsilon$  is suppressed for simplicity. By differentiation of the equation  $v = h(u)$  with respect to the independent variable  $t$ , it follows that  $F(h(u)) = Dh(u)f(u)$ . By linearization at the rest point  $u_0$ , we have

$$DF(h(u_0))Dh(u_0) = Dh(u_0)Df(u_0).$$

In particular the system matrix  $Df(h(u_0))$  for the linearized system at the perturbed rest point is similar, via the invertible matrix  $Dh(u_0)$ , to the system matrix  $Df(u_0)$  for the linearized system at the unperturbed rest point. Thus, the two rest points would have to have exactly the same eigenvalues, contrary to the general situation for which the eigenvalues of the perturbed linearization are different from the eigenvalues of the unperturbed linearization.

The notion of hyperbolicity is an essential hypothesis in structural stability theory; indeed, hyperbolic rest points are structurally stable. Likewise hyperbolic periodic orbits, defined to be those periodic orbits such that there is an associated hyperbolic Poincaré map are structurally stable. This notion can be extended to general invariant sets. The essential requirement is that all of the solutions of the linearized differential equations along the solutions in the invariant set, which are not initially in the direction of the invariant set, either grow or decay exponentially with uniform exponential growth rates over the entire invariant set.

The global theory of structural stability requires a global hypothesis to ensure that the rest points and period orbits are connected with orbits in general position.

**Definition 29** Let  $A$  be an invariant set (a union of orbits) of an autonomous first-order differential equation, vector field, or discrete dynamical system. The stable manifold of  $A$ , denoted  $W^s(A)$  is the set of all solutions that converge to  $A$  as  $t \rightarrow \infty$ . The unstable manifold  $W^u(A)$  is the set of all solutions that converge to  $A$  as  $t \rightarrow -\infty$ .

The stable manifold theorem asserts that the stable and unstable manifolds for rest points and periodic orbits of smooth vector fields are indeed (immersed) smooth manifolds (see, for example, [37,38]).

Two immersed manifolds are called transversal if they do not intersect; or, if the sum of their tangent spaces at each point of their intersection is the tangent space of the ambient manifold.

While periodicity is the most familiar form of recurrence for dynamical systems, a more subtle notion of recurrence is required for structural stability theorems.

**Definition 30** The point  $u$  is called a nonwandering point for an autonomous differential equation if, for each neighborhood  $U$  of  $u$  and each time  $t_0 > 0$ , there is some time  $t > t_0$  such that at least one solution starting in  $U$  at time zero returns to  $U$  at time  $t$ . Likewise for a discrete dynamical system defined by a diffeomorphism  $h$ , the point  $u$  is nonwandering if for each neighborhood  $U$  of  $u$  there is some integer  $k > 0$  such that  $h^k(U) \cap U$  is not empty. The set of all nonwandering points is called the nonwandering set.

For vector fields on compact orientable two-dimensional manifolds, the classical structural stability theorem is due to M. Peixoto [70] and is a generalization of earlier work for vector fields in the plane by L. Pointryagin and A. Andronov [2,3].

**Theorem 31** A class  $C^1$  vector field on a compact orientable two-dimensional manifold  $M$  is structurally stable if and only if

- (1) all rest points are hyperbolic,
- (2) all periodic solutions are hyperbolic,
- (3) stable and unstable manifolds of all pairs of saddle points (rest points whose linearizations have one positive and one negative eigenvalue) are disjoint,
- (4) the nonwandering set consists of only rest points and periodic solutions.

Moreover, the structurally stable vector fields form an open and dense set in  $\mathcal{X}(M)$ .

The notion of transversality in Peixoto's theorem is embodied in the property (3). The generalization of this property is an essential ingredient of the theory.

**Definition 32** A dynamical system satisfies the strong transversality property if for every pair of points  $\{u, v\}$  in its nonwandering set, the corresponding invariant manifolds  $W^s(u)$  and  $W^u(v)$  intersect transversally (at all their points of intersection).

**Definition 33** A dynamical system satisfies Axiom A if its nonwandering set is hyperbolic and the periodic orbits are dense in this set.

The most important result of the theory is the structural stability theorem:

**Theorem 34** *A class  $C^1$  diffeomorphism is structurally stable if and only if it satisfies Axiom A and the strong transversality property.*

Joel Robbin [77] proved that a  $C^2$  diffeomorphism is  $C^1$  structurally stable if it satisfies Axiom A and the strong transversality property. This result was improved by Clark Robinson [81] by relaxing the  $C^2$  requirement to  $C^1$ ; he also proved the analogous result for differential equations [79,80]. Ricardo Mañé [65] proved the converse.

A simple example of a structurally stable diffeomorphism is the north-pole south-pole map of the Riemann sphere; it is given by  $z \mapsto 2z$  for the complex variable  $z$ . There are two hyperbolic rest points, an unstable fixed point at  $z = 0$  and a stable fixed point at  $z = \infty$ . All other points are attracted to  $\infty$  under forward iteration and to zero under backward (inverse) iteration.

A more exotic example is provided by the hyperbolic toral automorphisms. To define these maps, note that a  $2 \times 2$ -integer matrix with determinant one determines a diffeomorphism of the plane that preserves the integer lattice. Thus, it projects to a map of the torus formed by the quotient space  $\mathbb{R}^2/\mathbb{Z}^2$ . In case the matrix is hyperbolic, the corresponding diffeomorphism on the torus has a dense set of periodic orbits and thus its nonwandering set is the entire torus. It can be proved that the linear hyperbolic structure of the transformation of the plane (that is, one stable and one unstable eigenspace) projects to a uniform hyperbolic structure on the entire torus; and, moreover, the strong transversality condition is satisfied. Thus, small perturbations of such a map have the same properties. This example was generalized by Dmitri Anosov and others to include a class of diffeomorphisms and vector fields in the continuous case now called Anosov dynamical systems. The most important example of a continuous flow with this property is the geodesic flow in the unit tangent bundle of a compact Riemannian manifold with negative sectional curvatures (see [4]).

## Attractors

A natural outgrowth of the theory of stability for rest points and periodic solutions is its generalization to invariant sets. The fundamental objects of study are the attractors. Unfortunately, the definition of this concept is not universally accepted; different authors use different definitions. The following definition is perhaps the most popular:

**Definition 35** A closed invariant set  $A$  for a dynamical system is called an attracting set if it is contained in an open neighborhood  $U$  of the ambient space such that the

intersection of the sequence of all forward motions of  $U$  under the action of the dynamical system is equal to  $A$ . An attracting set is called an attractor if, in addition, there is no closed invariant attracting subset of  $A$ .

The existence of attractors is part of the stability theory of dynamical systems; the structure of attractors is the subject of another rich theory. For most of the history of differential equations the only known attractors (the classical attractors) were rest points, periodic orbits, and tori, which are all manifolds. The existence of attractors with fractal dimensions, called strange attractors, was noticed and proved only recently [62,82,85,95].

For differential equations in  $\mathbb{R}^n$ , the existence of an attractor for autonomous systems is usually proved by demonstrating the existence of an  $(n-1)$ -dimensional sphere (or other separating hypersurface) such that the vector field corresponding to the differential equation points into the bounded region of space bounded by the sphere (more precisely, the inner product of the vector field and the inner normal is positive at all points on the sphere).

As an example, consider the Lorenz equations on  $\mathbb{R}^3$

$$\begin{aligned}\dot{x} &= \sigma(y - x), \\ \dot{y} &= rx - y - xz, \\ \dot{z} &= xy - bz,\end{aligned}$$

where  $\sigma$ ,  $r$ , and  $b$  are positive constants. The origin is a rest point for the Lorenz system and, if  $r < 1$ , then

$$L(x, y, z) := \frac{1}{\sigma}x^2 + y^2 + z^2$$

is a Lyapunov function. In this case, the origin is globally asymptotically stable. More generally (that is, with no restriction on the size of  $r$ ), there is a constant  $c > 0$  such that the vector field points into the bounded region of space bounded by the ellipsoid  $rx^2 + \sigma y^2 + \sigma(z - 2r)^2 = c$ . Thus, the Lorenz equations have an attractor. For some parameter values, for example  $\sigma = 10$ ,  $b = 8/3$ , and  $r = 28$ , the attractor is a fractal, called the Lorenz attractor (see [29,62,93,95]).

For the special case of two-dimensional first-order autonomous systems, the existence and nature of attractors can be more precisely specified. The main result is the Poincaré-Bendixson theorem (see [15] and, for a proof, [1]):

**Theorem 36** *An attractor of the class  $C^1$  differential Eq. (23) on the plane that contains no rest points is a periodic orbit.*



**Corollary 37** *If the differential Eq. (23) on the plane is analytic and has a positively invariant annulus containing no rest points, then the annulus contains an orbitally asymptotically stable periodic solution (a stable limit cycle).*

As an example, consider the model of a damped pendulum with torque given by the first-order planar system

$$\dot{\theta} = \nu, \quad \dot{\nu} = -\sin \theta + \mu - \lambda \nu, \quad (31)$$

where  $\lambda$  and  $\mu$  are positive constants. By viewing the variable  $\theta$  as an angular variable modulo  $2\pi$ , the phase space is the cylinder rather than the plane. Nonetheless, the Poincaré-Bendixson theory is valid on the cylinder because a bounded region of the cylinder can be flattened by a change of variables to a region of the plane. If  $|\mu| > 1$ , then the system of differential Eq. (31) has a globally attracting periodic orbit. Three main observations can be used to construct a proof: there are no rest points, the quantity  $-\sin \theta + \mu - \lambda \nu$  is negative for sufficiently large values of  $\nu > 0$ , and it is positive for negative values of  $\nu$  with sufficiently large absolute values. These facts imply the existence of an invariant annulus containing no rest points. The remainder of the proof uses Theorem 19 and the analyticity of solutions (see p. 98 in [19]).

### Generalizations and Future Directions

The concepts of stability theory have been successfully generalized to infinite-dimensional dynamical systems (which arise from the analysis of partial differential equations) defined on Banach spaces and to abstract dynamical systems on metric spaces [13,32,33,36,83,94]. Research in this direction has produced some results on stability and the existence of attractors.

No general theorem akin to Lyapunov's Theorem 4 is known for infinite-dimensional dynamical systems. A main difficulty is that, for an unbounded operator  $A$ , the non-zero elements in the spectrum of  $e^A$  may not all be given by the exponentials of elements in the spectrum of  $A$  (see, for a discussion, Chap. 2 in [20]). Thus, to determine the stability of a rest point by linearization, this spectral mapping property must be checked separately for most cases of interest.

Part of the motivation for the determination of attractors in infinite-dimensional dynamical systems is the possibility of reducing the long-term dynamical properties to the analysis of a finite-dimensional dynamical system on the attractor. The fundamental organizing concept in this direction is the "inertial manifold".

**Definition 38** An inertial manifold is a finite-dimensional Lipschitz manifold that is invariant under forward

motions of the dynamical system and attracts all solutions of the dynamical system at an exponential rate.

Inertial manifolds have been proved to exist for many dynamical systems defined by evolution-type partial differential equations, for example, reaction-diffusion equations with appropriate boundary conditions. Much current research is devoted to determining conditions that imply the existence of inertial manifolds (or their generalizations) for the equations of fluid dynamics.

The conceptual framework and basic abstract theory of stability is mature and well-understood for finite-dimensional dynamical systems. This theory is currently being extended to an abstract theory for infinite-dimensional dynamical systems. The most important direction for further development is the application of this theory to specific differential equations (including partial differential equations) that arise as mathematical models in applied mathematics. Even the problem that originally motivated stability theory (that is, the stability of periodic and quasi-periodic motions of the  $N$ -body problem of classical mechanics) remains open to future research.

### Bibliography

#### Primary Literature

1. Alligood KT, Sauer TD, Yorke JA (1997) *Chaos: An introduction to dynamical systems*. Springer, New York
2. Andronov AA, Vitt EA, Khaiken SE (1966) *Theory of Oscillations*. Pergamon Press, Oxford
3. Andronov AA, Leontovich EA, Gordon II, Maier AG (1973) *Qualitative Theory of Second-Order Dynamic Systems*. Wiley, New York
4. Anosov DV (1967) Geodesic Flows on Closed Riemannian Manifolds with Negative Curvature. *Proc Steklov Inst Math* 90:1–209
5. Arnold VI (1963) Proof of an Kolmogorov's theorem on the preservation of quasi-periodic motions under small perturbations of the Hamiltonian. *Usp Mat Nauk SSSR* 113(5):13–40
6. Arnold VI (1964) Instability of dynamical systems with many degrees of freedom. *Soviet Math* 5(3):581–585
7. Arnold VI (1973) *Ordinary Differential Equations*. MIT Press, Cambridge
8. Arnold VI (1978) *Mathematical Methods of Celestial Mechanics*. Springer, New York
9. Arnold VI (1982) *Geometric Methods In: The Theory of Ordinary Differential Equations*. Springer, New York
10. Arnold VI (ed) (1988) *Dynamical Systems I*. Springer, New York
11. Arnold VI (ed) (1988) *Dynamical Systems III*. Springer, New York
12. Arscott FM (1964) *Periodic Differential Equations*. MacMillan, New York
13. Babin AV, Vishik MI (1992) *Attractors of Evolution Equations*. North-Holland, Amsterdam

14. Bautin NN (1954) On the number of limit cycles which appear with the variation of coefficients from an equilibrium position of focus or center type. *Am Math Soc Transl* 100:1–19
15. Bendixson I (1901) Sur les courbes définies par des équations différentielles. *Acta Math* 24:1–88
16. Brjuno AD (1971) Analytic form of differential equations. *Trudy MMO* 25:119–262
17. Chang K-C (2005) *Methods in Nonlinear Analysis*. Springer, New York
18. Chetayev NG (1961) *The Stability of Motion*. Pergamon Press, New York
19. Chicone C (2006) *Ordinary Differential Equations with Applications*, 2nd edn. Springer, New York
20. Chicone C, Latushkin Y (1999) Evolution Semigroups. In: *Dynamical Systems and Differential Equations*. Am Math Soc, Providence
21. Chicone C, Liu W (2004) Asymptotic phase revisited. *J Diff Eq* 204(1):227–246
22. Chicone C, Swanson R (2000) Linearization via the Lie derivative. *Electron J Diff Eq Monograph* 02:64
23. Dirichlet GL (1846) Über die Stabilität des Gleichgewichts. *J Reine Angewandte Math* 32:85–88
24. Dirichlet GL (1847) Note sur la stabilité de l'équilibre. *J Math Pures Appl* 12:474–478
25. Dulac H (1908) Détermination et intégration d'une certaine classe d'équations différentielles ayant pour point singulier un centre. *Bull Soc Math Fr* 32(2):230–252
26. Dumortier F (2006) Asymptotic phase and invariant foliations near periodic orbits. *Proc Am Math Soc* 134(10):2989–2996
27. Floquet G (1883) Sur les équations différentielles linéaires à coefficients périodiques. *Ann Sci École Norm Sup* 12(2):47–88
28. Frommer M (1934) Über das Auftreten von Wirbeln und Strudeln (geschlossener und spiralförmiger Integralkurven) in der Umgebung rationaler Unbestimmtheitsstellen. *Math Ann* 109:395–424
29. Guckenheimer J, Holmes P (1986) *Nonlinear Oscillations, Dynamical Systems, and Bifurcation of Vector Fields*, 2nd edn. Springer, New York
30. Grobman D (1959) Homeomorphisms of systems of differential equations. *Dokl Akad Nauk SSSR* 128:880–881
31. Grobman D (1962) Topological classification of the neighborhood of a singular point in  $n$ -dimensional space. *Mat Sb* 98:77–94
32. Hale JK (1988) *Asymptotic Behavior of Dissipative Systems*. Am Math Soc, Providence
33. Hale JK, Magalhães LT, Oliva WM (2002) *Dynamics in Infinite Dimensions*, 2nd edn. Springer, New York
34. Hartman P (1960) A lemma in the theory of structural stability of differential equations. *Proc Am Math Soc* 11:610–620
35. Hartman P (1960) On local homeomorphisms of Euclidean space. *Bol Soc Math Mex* 5(2):220–241
36. Henry D (1981) *Geometric Theory of Semilinear Parabolic Equations*. Lecture Notes in Mathematics, vol 840. Springer, New York
37. Hirsch M, Pugh C (1970) Stable manifolds and hyperbolic sets. In: *Global Analysis XIV*. Am Math Soc, Providence, pp 133–164
38. Hirsch MC, Pugh C, Shub M (1977) *Invariant Manifolds*. Lecture Notes in Mathematics, vol 583. Springer, New York
39. Hofer H, Zehnder E (1994) *Symplectic invariants and Hamiltonian dynamics*. Birkhäuser, Basel
40. Hurwitz A (1895) Über die Bedingungen, unter welchen eine Gleichung nur Wurzeln mit negativen reellen Theilen besitzt. [On the conditions under which an equation has only roots with negative real parts.] *Math Ann* 46:273–284; (1964) In: Ballman RT et al (ed) *Selected Papers on Math Trends in Control Theory*. Dover, New York (reprinted)
41. Ilyashenko JS (1972) Algebraic unsolvability and almost algebraic solvability of the problem for the center-focus. *Funkcional Anal i Prilozhen* 6(3):30–37
42. Jost R (1968) Winkel- und Wirkungsvariable für allgemeine mechanische Systeme. *Helvetica Phys Acta* 41:965–968
43. Kapteyn W (1911) On the centre of the integral curves which satisfy differential equations of the first order and the first degree. *Proc Kon Akad Wet Amsterdam* 13:1241–1252
44. Kapteyn W (1912) New researches upon the centre of the integrals which satisfy differential equations of the first order and the first degree. *Proc Kon Akad Wet Amsterdam* 14: 1185
45. Kapteyn W (1912) New researches upon the centre of the integrals which satisfy differential equations of the first order and the first degree. *Proc Kon Akad Wet Amsterdam* 15:46–52
46. Kolmogorov AN (1954) La théorie générale des systèmes dynamiques et la mécanique. *Amsterdam Congress I*, pp 315–333
47. Kolmogorov AN (1954) On the conservation of conditionally periodic motions under small perturbations of the Hamiltonian. *Dokl Akad Nauk SSSR* 98:527–530
48. Lagrange JL (1776) *Nouv Mem Acad Roy Sci. Belles-Lettres de Berlin*, p 199
49. Lagrange JL (1788) *Mechanique Analytique*. Desaint, Paris, pp 66–73, pp 389–427; (1997) *Boston Studies in the Philosophy of Science* 191. Kluwer, Dordrecht (reprinted)
50. Lagrange JL (1869) *Oeuvres de Lagrange*. Gauthier, Paris 4:255
51. LaSalle JP (1976) *The Stability of Dynamical Systems*. SIAM, Philadelphia
52. Laplace PS (1784) *Mem Acad Roy Sci. Paris*, p 1
53. Laplace PS (1895) *Oeuvres Complètes de Laplace*. Gauthier, Paris 11:47
54. Leontovich AM (1962) On the stability of the Lagrange periodic solutions for the reduced problem of three bodies. *Sov Math Dokl* 3(2):425–430
55. Liapounoff AM (1947) *Problème général de la stabilité du mouvement*. Princeton University Press, Princeton
56. Liouville J (1842) Extrait d'une lettre à M Arago sur le Mémoire de M Maurice relatif à l'invariabilité des grands axes des orbites planétaires. *Compt Rend Séances l'Acad Sci* 15:425–426
57. Liouville J (1842) Extrait d'un Mémoire sur un cas particulier du problème des trois corps. *J Math Sér I* 7:110–113
58. Liouville J (1843) Sur la loi de la pesanteur à la surface ellipsoïdale d'équilibre d'une masse liquide homogène douée d'un mouvement de rotation. *J Math Sér I* 8:360
59. Liouville J (1855) Formules générales relatives à la question de la stabilité de l'équilibre d'une masse liquide homogène douée d'un mouvement de rotation autour d'un axe. *Sér I, J Math Pures Appl* 20:164–184
60. Lochak P (1999) Arnold diffusion; a compendium of remarks and questions. *Hamiltonian systems with three or more degrees of freedom* (S'Agaró 1995). In: *NATO Adv Sci Inst Ser C Math Phys Sci*, 533. Kluwer, Dordrecht, pp 168–183

61. Lochak P, Meunier C (1988) Multiphase averaging for classical systems. In: Trans H.S. Dumas. Springer, New York
62. Lorenz EN (1963) Deterministic nonperiodic flow. *J Atm Sci* 20:130–141
63. Lyapunov AM (1892) The General Problem of the Stability of Motion. *Russ Math Soc, Kharkov* (russian); (1907) *Ann Facult Sci l'Univ Toulouse* 9:203–474 (french trans: Davaux É); (1949) Princeton University Press, Princeton (reprinted); (1992) Taylor & Francis, London (english trans: Fuller AT)
64. Magnus W, Winkler S (1979) Hill's Equation. Dover, New York
65. Mañé R (1988) A proof of the  $C^1$  stability conjecture. *Inst Hautes Études Sci Publ Math* 66:161–210
66. Marcus L, Yamabe H (1960) Global stability criteria for differential equations. *Osaka Math J* 12:305–317
67. Melnikov VK (1963) On the stability of the center for time periodic perturbations. *Trans Mosc Math Soc* 12:1–57
68. Moser J (1962) On invariant curves of area-preserving mappings of an annulus. *Nachr Akad Wiss Göttingen, Math Phys* K1:1–10
69. Moser J (1978/79) Is the solar system stable? *Math Intell* 1: 65–71
70. Peixoto MM (1962) Structural stability on two-dimensional manifolds. *Topology* 1:101–120
71. Poincaré H (1881) Mémoire sur les courbes définies par une équation différentielle. *J Math Pures Appl* 7(3):375–422; (1882) 8:251–296; (1885) 1(4):167–244; (1886) 2:151–217; all reprinted (1928) Tome I. Gauthier-Villars, Oeuvre, Paris
72. Poincaré H (1890) Sur le problème des trois corps et les équations de la dynamique. *Acta Math* 13:83–97
73. Poincaré H (1892) Les méthodes nouvelles de la mécanique céleste. Tome III. Invariants intégraux Solutions périodiques du deuxième genre, Solutions doublement asymptotiques. Dover, New York (1957)
74. Poincaré H (1951) Oeuvres, vol 1. Gauthier-Villars, Paris
75. Poincaré H (1957) Les méthodes nouvelles de la mécanique céleste. Tome I. Solutions périodiques. Non-existence des intégrales uniformes, Solutions asymptotiques. Dover, New York
76. Poincaré H (1957) Les méthodes nouvelles de la mécanique céleste. Tome II. Méthodes de MM. Newcomb, Gylden, Lindstedt et Bohlin. Dover, New York
77. Robbin JW (1971) A structural stability theorem. *Ann Math* 94(2):447–493
78. Robbin JW (1995) Matrix algebra. A.K. Peters, Wellesley
79. Robinson C (1974) Structural stability of vector fields. *Ann Math* 99(2):154–175
80. Robinson C (1975) Errata to: "Structural stability of vector fields". (1974) *Ann Math* 99(2):154–175, *Ann Math* 101(2):368
81. Robinson C (1976) Structural Stability for  $C^1$  diffeomorphisms. *J Diff Eq* 22(1):28–73
82. Robinson C (1989) Homoclinic bifurcation to a transitive attractor of Lorenz type. *Nonlinearity* 2(4):495–518
83. Robinson J (2001) Infinite Dimensional Dynamical Systems. Cambridge Univ Press, Cambridge
84. Routh EJ (1877) A Treatise on the Stability of a Given State of Motion. Macmillan, London; Fuller AT (ed) (1975) Stability of Motion. Taylor, London (reprinted)
85. Rychlik MR (1990) Lorenz attractors through Šil'nikov-type bifurcation I. *Ergodic Theory Dynam Syst* 10(4):793–821
86. Sanders JA, Verhulst F, Murdock J (2007) Averaging Methods. In: Nonlinear Dynamical Systems, 2nd edn. Appl Math Sci 59. Springer, New York
87. Siegel CL (1942) Iterations of analytic functions. *Ann Math* 43:607–612
88. Siegel CL (1952) Über die Normalform analytischer Differentialgleichungen in der Nähe einer Gleichgewichtslösung. *Nachr Akad Wiss Göttingen, Math Phys* K1:21–30
89. Siegel CL, Moser J (1971) Lectures on Celestial Mechanics. Springer, New York
90. Sternberg S (1958) On the structure of local homeomorphisms of Euclidean  $n$ -space. *Am J Math* 80(3):623–631
91. Sternberg S (1959) On the structure of local homeomorphisms of Euclidean  $n$ -space. *Am J Math* 81(3):578–604
92. Sternberg S (1969) Celestial Mechanics. Benjamin, New York
93. Strogatz SH (1994) Nonlinear Dynamics and Chaos. Perseus Books, Cambridge
94. Temam R (1997) Infinite-Dimensional Dynamical Systems. In: Mechanics and Physics. Springer, New York
95. Viana M (2000) What's new on Lorenz strange attractors? *Math Intell* 22(3):6–19
96. Wiggins S (2003) Introduction to Applied Nonlinear Dynamical Systems and Chaos, 2nd edn. Springer, New York
97. Willems JL (1970) Stability Theory of Dynamical Systems. Wiley, New York
98. Yakubovich VA, Starzhinskii VM, Louvish D (trans) (1975) Linear Differential Equations with Periodic Coefficients. Wiley, New York
99. Żołądek H (1994) Quadratic systems with center and their perturbations. *J Diff Eq* 109(2):223–273
100. Żołądek H (1997) The problem of the center for resonant singular points of polynomial vector fields. *J Diff Eq* 135:94–118

### Books and Reviews

- Arnold VI, Avez A (1968) Ergodic Problems of Classical Mechanics. Benjamin, New York
- Coddington EA, Levinson N (1955) Theory of ordinary differential equations. McGraw-Hill, New York
- Farkas M (1994) Periodic Motions. Springer, New York
- Glendinning PA (1994) Stability, Instability and Chaos. Cambridge University Press, Cambridge
- Grimshaw R (1990) Nonlinear Ordinary Differential Equations. Blackwell, Oxford
- Hale JK (1980) Ordinary Differential Equations, 2nd edn. RE Krieger, Malabar
- Hahn W, Baartz AP (trans) (1967) Stability of Motion. Springer, New York
- Hartman P (1964) Ordinary Differential Equations. Wiley, New York
- Hirsch M, Smale S (1974) Differential Equations, Dynamical Systems, and Linear Algebra. Acad Press, New York
- Katok AB, Hasselblatt B (1995) Introduction to the Modern Theory of Dynamical Systems. Cambridge Univ Press, Cambridge
- Krasovskii NN (1963) Stability of Motion. Stanford Univ Press, Stanford
- LaSalle J, Lefschetz S (1961) Stability by Lyapunov's Direct Method with Applications. Acad Press, New York
- Lefschetz S (1977) Differential Equations: Geometric Theory, 2nd edn. Dover, New York
- Miller RK, Michel AN (1982) Ordinary Differential Equations. Acad Press, New York

- Moser J (1973) Stable and Random Motions. In: Dynamical Systems Ann Math Studies 77. Princeton Univ Press, Princeton
- Nemytskii VV, Stepanov VV (1960) Qualitative Theory of Differential Equations. Princeton Univ Press, Princeton
- Perko L (1996) Differential Equations and Dynamical Systems, 2nd edn. Springer, New York
- Robbin JW (1970) On structural stability. Bull Am Math Soc 76:723–726
- Robbin JW (1972) Topological conjugacy and structural stability for discrete dynamical systems. Bull Am Math Soc 78(6):923–952
- Robinson C (1995) Dynamical Systems: Stability, Symbolic Dynamics, and Chaos. CRC Press, Boca Raton
- Smale S (1967) Differentiable dynamical systems. Bull Am Math Soc 73:747–817
- Smale S (1980) The Mathematics of Time: Essays on Dynamical Systems, Economic Processes and Related Topics. Springer, New York
- Sotomayor J (1979) Lições de equações diferenciais ordinárias. IMPA, Rio de Janeiro
- Verhulst F (1989) Nonlinear Differential Equations and Dynamical Systems. Springer, New York

## Static Games

OSCAR VOLIJ

Ben-Gurion University, Beer-Sheva, Israel

### Article Outline

Glossary

Definition of the Subject

Introduction

Nash Equilibrium

Existence

Mixed Strategies

Equilibrium in Beliefs

Correlated Equilibrium

Rationality, Correlated Equilibrium  
and Equilibrium in Beliefs

Bayesian Games

Evolutionary Stable Strategies

Future Directions

Bibliography

### Glossary

**Player** A participant in a game.

**Action set** The set of actions that a player may choose.

**Action profile** A list of actions, one for each player.

**Payoff** The utility a player obtains from a given action profile.

### Definition of the Subject

Game theory concerns the interaction of decision makers. This interaction is modeled by means of games. There are

various approaches to constructing games. One approach is to focus on the possible outcomes of the decision-makers' interaction by abstracting from the actions or decisions that may lead to these outcomes. The main tool used to implement this approach is the *cooperative game*. Another approach is to focus on the actions that the decision-makers can take, the main tool being the *non-cooperative game*. Within this approach, strategic interactions are modeled in two ways. One is by means of *dynamic*, or extensive form games, and the other is by means of *static*, or strategic games. Dynamic games stress the sequentiality of the various decisions that agents can make. An essential component of a dynamic game is the description of who moves first, who moves second, etc. Static games, on the other hand, abstract from the sequentiality of the possible moves, and model interactions as simultaneous decisions, where the decisions may well be complicated plans of actions that dictate different moves for different situations that may arise. All extensive form games can be modeled as static games, and all strategic form games can be modeled as extensive form games. But some situations may be more conveniently modeled as one or the other kind of game.

This chapter reviews the main ideas and results related to static games, as well as some interesting relationships that connect equilibrium concepts with the idea of rationality. The objective is to introduce the reader to the area of static games and to stimulate his interest for further knowledge of game theory in general. For a comprehensive exposition of some results not covered in this chapter, the reader is referred to the many excellent textbooks available on game theory. Binmore [6], Fudenberg and Tirole [9], Osborne [20], Osborne and Rubinstein [21] constitute only a partial list.

Although the definition of a static game is a very simple one, static games are a very flexible model which allows us to analyze many different situations. In particular, one can use them to analyze strategic interactions that involve either common interests or diametrically opposed interests. Similarly, one can also use static games to model situations where players have either symmetric or asymmetric information. The range of applications of static games is very wide and covers many disciplines, such as economics, political science, biology, philosophy, and computer science among others.

### Introduction

In this section we introduce some examples that will be used later to motivate different concepts. We also introduce the definition of a static game.



**The prisoner's dilemma** involves a donor who is interested in donating some amount of money to two universities. The donor decides that the amount each university will receive depends on the content of the messages the presidents of the respective universities will send to him. Each university will send simultaneously one of two messages. One possible message is "Give him 2" and the other is "Give me 1". The donor will do exactly as told. For instance, if University I sends the message "Give me 1" and University II sends "Give him 2", the donor will donate \$3 to University I and \$0 to University II. This game can be described by means of the following matrix, where the entries represent the payoffs for University I and University II, respectively, that result from the corresponding action choices.

|              |            | University II |           |
|--------------|------------|---------------|-----------|
|              |            | Give him 2    | Give me 1 |
| University I | Give him 2 | 2, 2          | 0, 3      |
|              | Give me 1  | 3, 0          | 1, 1      |

**The battle of the sexes** consists of two friends, She and He, who want to go out together, but have no means of communication. They have to decide, each one separately but both simultaneously, whether to go to a boxing match or to a ballet show. For both of them, the worst possible outcome would be to choose different events and not meet. But if they meet, he would rather meet her at the boxing match, while she would rather meet him at the ballet. The battle of the sexes can be described by the following matrix.

|    |        | She  |        |
|----|--------|------|--------|
|    |        | Box  | Ballet |
| He | Box    | 2, 1 | 0, 0   |
|    | Ballet | 0, 0 | 1, 2   |

Again, the entries of this matrix represent the payoffs that he and she get, as a result of their corresponding choices.

**Chicken** models two drivers who approach each other on a narrow street. If none of them slows down they'll have an accident and their corresponding payoffs will be 0. But if at least one of them slows down, the accident is prevented. The problem is that both of them would like the other to slow down. If only one driver slows down, this driver gets a payoff of 2 and the other driver gets a payoff of 7. If both drivers slow down, then both drivers get a payoff of 6. This situation can

be described by the following matrix.

|          |           | Driver 2  |          |
|----------|-----------|-----------|----------|
|          |           | Slow Down | Speed up |
| Driver 1 | Slow Down | 6, 6      | 2, 7     |
|          | Speed up  | 7, 2      | 0, 0     |

**Matching Pennies** involves two friends, each of whom places a coin on a table. If both coins are placed heads up or tails up, then friend 1 gets one dollar from friend 2. If one coin is placed heads up and the other tails up, then friend 1 pays one dollar to friend 2. Matching pennies can be described by the following matrix, where the entries are the amounts of money that the friends get from each other.

|          |       | Friend 2 |       |
|----------|-------|----------|-------|
|          |       | Heads    | Tails |
| Friend 1 | Heads | 1, -1    | -1, 1 |
|          | Tails | -1, 1    | 1, -1 |

The above examples of strategic interactions can be modeled as static games. A static game is a formalization of a strategic situation according to which players choose their actions separately and simultaneously, and as a result obtain certain payoffs. The interaction that a static game models need not require that players take their actions simultaneously. But the interaction is modeled by defining actions in such a way that lets us think of the players as acting simultaneously.

All of the above examples involve a set of players, and for each player there is a set of available actions and a function that associates a payoff level to each of the profiles of actions that may result from the players' choices. These are the three essential components of a static game, as formalized in the following definition.

**Definition 1** A static game is a triple  $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  where  $N$  is a finite set of players, and for each player  $i \in N$ ,  $A_i$  is  $i$ 's set of actions, and  $u_i: \times_{k \in N} A_k \rightarrow \mathbb{R}$  is player  $i$ 's utility function.

In the prisoner's dilemma the set of players is  $N = \{\text{University I, University II}\}$ ; the sets of actions are  $A_I = \{\text{Give me 1, Give him 2}\}$ ; the utility function of University I is  $u_I(\text{Give me 1, Give me 1}) = 1$ ,  $u_I(\text{Give me 1, Give him 2}) = 3$ ,  $u_I(\text{Give him 2, Give me 1}) = 0$ ,  $u_I(\text{Give him 2, Give him 2}) = 2$ ; and the utility function of University II is  $u_{II}(\text{Give me 1, Give me 1}) = 1$ ,  $u_{II}(\text{Give me 1, Give him 2}) = 0$ ,  $u_{II}(\text{Give him 2, Give me 1}) = 3$ ,  $u_{II}(\text{Give him 2, Give him 2}) = 6$ .

In this chapter we sometimes refer to static games simply as games. For any game  $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ , the set



of action profiles  $\times_{k \in N} A_k$  is denoted by  $A$ , and a typical action profile is denoted by  $a = (a_i)_{i \in N} \in A$ . If  $A$  is a finite set, then we say that the game is finite. Player  $i$ 's utility function represents his preferences over the set of action profiles. For instance, for any two action profiles  $a$  and  $a'$  in  $A$ ,  $u_i(a) \geq u_i(a')$  means that player  $i$  prefers action profile  $a$  to action profile  $a'$ . Clearly, although player  $i$  has preferences over action profiles, he can only affect his own component,  $a_i$ , of the profile.

### Nash Equilibrium

One objective of game theory is to select, for each game, a set of action profiles that are interesting in some way. These action profiles may be interpreted as predictions of the theory, or prescriptions for the players to follow, or simply as equilibrium outcomes in the sense that if they occur, the players do not wish that they had acted differently. These action profiles are formally given by solution concepts, which are functions that associate each strategic game with the selected set of action profiles. The central solution concept in game theory is known as *Nash equilibrium*. The hypothesis behind this solution concept is that each player chooses his actions so as to maximize his utility, given the profile of actions chosen by the other players. To give a formal definition of the Nash equilibrium concept, we first introduce some useful notation. For each player  $i \in N$ , let  $A_{-i} = \times_{k \in N \setminus \{i\}} A_k$  be the set of the other players' profiles of actions. Then we can write  $A = A_i \times A_{-i}$ , and each action profile can be written as  $a = (a_i, a_{-i}) \in A_i \times A_{-i}$ , thereby distinguishing player  $i$ 's action from the other players' profile of actions.

**Definition 2** The action profile  $a^* = (a_i^*)_{i \in N} \in A$  in a game  $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  is a *Nash equilibrium* if for each player,  $i \in N$ , and every action  $a_i \in A_i$  of player  $i$ ,  $a^*$  is at least as good for player  $i$  as the action profile  $(a_i, a_{-i}^*)$ . That is, if

$$u_i(a^*) \geq u_i(a_i, a_{-i}^*) \quad \text{for all } a_i \in A_i \quad \text{and for all } i \in N.$$

It is a *strict Nash equilibrium* if the above inequality is strict for all alternative actions  $a_i \in A_i \setminus \{a_i^*\}$ .

### Analysis of Some Finite Games

**Prisoner's Dilemma** Recall that the prisoner's dilemma can be described by the following matrix.

|              |            | University II |           |
|--------------|------------|---------------|-----------|
|              |            | Give him 2    | Give me 1 |
| University I | Give him 2 | 2, 2          | 0, 3      |
|              | Give me 1  | 3, 0          | 1, 1      |

The action profile (*Give me 1*, *Give me 1*) is a Nash equilibrium. Indeed,

$$\begin{aligned} u_I(\text{Give me 1, Give me 1}) &= 1 \\ &\geq u_I(\text{Give him 2, Give me 1}) = 0 \end{aligned}$$

and

$$\begin{aligned} u_{II}(\text{Give me 1, Give me 1}) &= 1 \\ &\geq u_{II}(\text{Give me 1, Give him 2}) = 0. \end{aligned}$$

On the other hand, the action profile (*Give him 2*, *Give him 2*) is not a Nash equilibrium, since University I prefers action "*Give me 1*" if University II chooses action "*Give him 2*":

$$\begin{aligned} 2 &= u_I(\text{Give him 2, Give him 2}) \\ &< u_I(\text{Give me 1, Give him 2}) = 3. \end{aligned}$$

**Battle of the Sexes** Recall that the battle of the sexes can be described by the following matrix.

|    |        | She  |        |
|----|--------|------|--------|
|    |        | Box  | Ballet |
| He | Box    | 2, 1 | 0, 0   |
|    | Ballet | 0, 0 | 1, 2   |

One can check that (*Box*, *Box*) is a Nash equilibrium and (*Ballet*, *Ballet*) is a Nash equilibrium as well. It can also be checked that these are the only two action profiles that constitute a Nash equilibrium.

**Matching Pennies** The reader can check that Matching Pennies has no Nash equilibrium.

Before we analyze the next example, we introduce a technical tool that allows us to reformulate the definition of Nash equilibrium more conveniently. More importantly, this alternative definition is the key to the standard proof of the existence of Nash equilibrium.

**Definition 3** Let  $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  be a strategic game and let  $i \in N$  be a player. Consider a list of actions  $a_{-i} = (a, \dots, a_{i-1}, a_{i+1}, \dots, a_n) \in \times_{k \in N \setminus \{i\}} A_k$  of all the players other than  $i$ . The set of player  $i$ 's best responses to  $a_{-i}$  is

$$\begin{aligned} \mathcal{B}_i(a_{-i}) &= \{a_i \in A_i : u_i(a_i, a_{-i}) \geq u_i(b_i, a_{-i}) \\ &\quad \text{for all } b_i \in A_i\}. \end{aligned}$$

The correspondence  $\mathcal{B}_i : \times_{j \neq i} A_j \rightrightarrows A_i$  that assigns to each  $(n-1)$ -tuple of actions in  $A_{-i}$  the set of best responses to it is called the *best response correspondence* of player  $i$ .

The definition of a Nash equilibrium may be stated in terms of the players' best response correspondences, as stated in the following proposition.

**Proposition 1** *The action profile  $a^* \in A$  is a Nash equilibrium if and only if every player's action is a best response to the other players' actions. That is, if*

$$a_i^* \in \mathcal{B}_i(a_{-i}^*) \quad \text{for all } i \in N.$$

Until now, all the examples involved games where the action sets contained two actions. The next example is a game where the players' action sets are infinite. We will use the player's best response correspondences to find all its Nash equilibria.

**The War of Attrition** Two animals, 1 and 2, are fighting over a prey. Each animal chooses a time at which it intends to give up. Once one animal has given up, the other obtains the prey; if both animals give up at the same time then they split the prey equally. For each  $i = 1, 2$ , animal  $i$ 's willingness to fight for the prey is given by  $v_i > 0$ . The value  $v_i$  is the maximum amount of time that animal  $i$  is willing to spend to obtain the prey. Since fighting is costly, each animal prefers as short a fight as possible. If animal  $i$  obtains the prey after a fight of length  $t$ , his utility will be  $v_i - t$ . We can model the situation as the game  $G = (\{1, 2\}, (A_1, A_2), (u_1, u_2))$  where

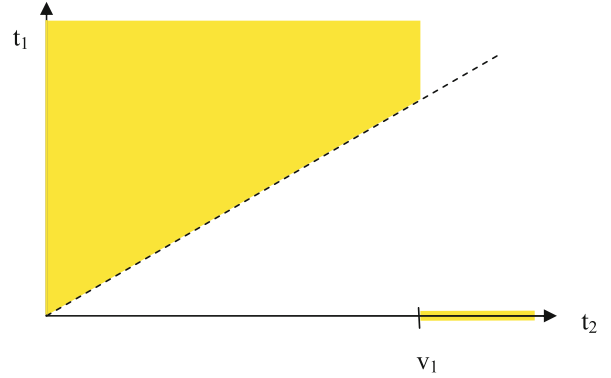
- $A_1 = [0, \infty] = A_2$  (an element  $t \in A_i$  represents a time at which player  $i$  plans to give up)

$$u_1(t_1, t_2) = \begin{cases} -t_1 & \text{if } t_1 < t_2 \\ \frac{1}{2}v_1 - t_2 & \text{if } t_1 = t_2 \\ v_1 - t_2 & \text{if } t_1 > t_2 \end{cases}$$

$$u_2(t_1, t_2) = \begin{cases} -t_2 & \text{if } t_2 < t_1 \\ \frac{1}{2}v_2 - t_1 & \text{if } t_1 = t_2 \\ v_2 - t_1 & \text{if } t_2 > t_1. \end{cases}$$

We are interested in the best response correspondences. First, we calculate player 1's best response correspondence,  $\mathcal{B}_1(t_2)$ . There are three cases to consider.

**Case 1:  $t_2 < v_1$**  In this case,  $v_1 - t_2 > \frac{1}{2}v_1 - t_2$  and  $v_1 - t_2 > -t_1$ . Consequently, given that player 2's action is  $t_2$ , player 1's utility function has a maximum value of  $v_1 - t_2$ , which is attained at any  $t_1 > t_2$ . Therefore,  $\mathcal{B}_1(t_2) = (t_2, \infty)$ .



Static Games, Figure 1

Player 1's best response correspondence

**Case 2:  $t_2 = v_1$**  In this case,  $0 = v_1 - t_2 > \frac{1}{2}v_1 - t_2$ .

Therefore, player's 1 utility function  $u_1(\cdot, t_2)$  has a maximum value of 0, which is attained at  $t_1 = 0$  and at  $t_1 > t_2$ . Therefore,  $\mathcal{B}_1(t_2) = \{0\} \cup (t_2, \infty)$ .

**Case 3:  $t_2 > v_1$**  In this case  $\frac{1}{2}v_1 - t_2 < v_1 - t_2 < 0$ .

As a result, player 1's utility function  $u_1(\cdot, t_2)$  has a maximum value of 0, which is attained at  $t_1 = 0$ . Therefore,  $\mathcal{B}_1(t_2) = \{0\}$ .

Summarizing, player 1's best response correspondence is:

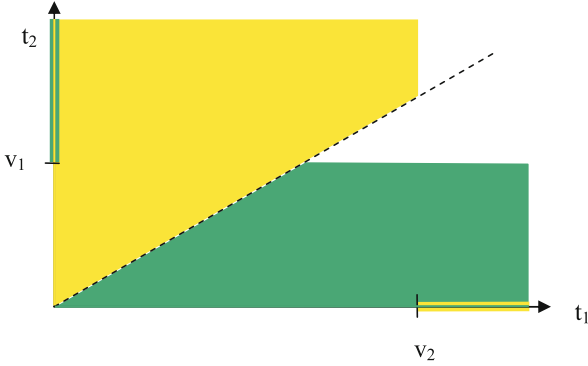
$$\mathcal{B}_1(t_2) = \begin{cases} (t_2, \infty) & \text{if } t_2 < v_1 \\ \{0\} \cup (t_2, \infty) & \text{if } t_2 = v_1 \\ \{0\} & \text{if } t_2 > v_1 \end{cases}$$

which is depicted in Fig. 1.

Similarly, player 2's best response correspondence is:

$$\mathcal{B}_2(t_1) = \begin{cases} (t_1, \infty) & \text{if } t_1 < v_2 \\ \{0\} \cup (t_1, \infty) & \text{if } t_1 = v_2 \\ \{0\} & \text{if } t_1 > v_2. \end{cases}$$

Combining the two best response correspondences we get that  $(t_1^*, t_2^*)$  is a Nash equilibrium if and only if either  $t_1^* = 0$  and  $t_2^* \geq v_1$  or  $t_2^* = 0$  and  $t_1^* \geq v_2$ . Figure 2 depicts the set of all the Nash equilibria as the intersection of the two best response correspondences. Two things are worth noting. First, it is not necessarily the case that the player who values the prey most wins the war. That is, there are Nash equilibria of the war of attrition where the player with the highest willingness to fight for the prey gives in first, and as a result the object goes to the other player. Second, in none of the Nash equilibria is there a physical fight. All Nash



Static Games, Figure 2  
The equilibria

equilibria involve one player giving in immediately to the other. This second feature seems rather unrealistic, since fights in “war of attrition”-like situations are commonly observed. If one wants to obtain a fight of positive length in the war of attrition one needs to either drop the Nash equilibrium concept and adopt an alternative one, or model the war of attrition differently. We will adopt this second course of action later.

### Existence

As the matching pennies example shows, not all games have a Nash equilibrium. The following theorem, which dates back to Nash [18] and Glicksberg [11], states sufficient conditions on a game for it to have a Nash equilibrium. An earlier version of this theorem for the smaller but prominent class of zero-sum games can be found in von Neumann [23] (translated in von Neumann [24]). The standard proofs use Kakutani’s fixed point theorem. We present here an alternative proof, due to Geanakoplos [10], which uses Brouwer’s fixed point theorem instead.

**Theorem 1** *The game  $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  has a Nash equilibrium if for all  $i \in N$*

- *the set  $A_i$  of actions of player  $i$  is a nonempty compact convex subset of an Euclidean space,*
- *the utility function  $u_i$  is continuous,*
- *the utility function  $u_i$  is concave in  $A_i$ .*

*Proof (Geanakoplos)* Define the correspondence  $\varphi_i: A \rightrightarrows A_i$  by

$$\varphi_i(\bar{a}) = \arg \max_{a_i \in A_i} \{U_i(a_i, \bar{a}_{-i}) - \|a_i - \bar{a}_i\|^2\},$$

where,  $\|\cdot\|$  denotes a norm in the relevant Euclidean space. Note first that  $\varphi_i$  is a nonempty valued correspon-

dence because the maximand is a continuous function and  $A_i$  is compact. Second, note that the function  $\|a_i - \bar{a}_i\|$  is convex:

$$\begin{aligned} & \|(\lambda a_i + (1-\lambda)b_i) - \bar{a}_i\| \\ &= \|(\lambda a_i - \lambda \bar{a}_i) + ((1-\lambda)b_i - (1-\lambda)\bar{a}_i)\| \\ &\leq \|(\lambda a_i - \lambda \bar{a}_i)\| + \|((1-\lambda)b_i - (1-\lambda)\bar{a}_i)\| \\ &\leq |\lambda| \|a_i - \bar{a}_i\| + |1-\lambda| \|b_i - \bar{a}_i\|. \end{aligned}$$

Since the quadratic function is strictly convex, then the maximand is a strictly concave function. Therefore, the correspondence  $\varphi_i$  is in fact a function. Furthermore, since the maximand is continuous in the parameter  $\bar{a}$ ,  $\varphi_i$  is also continuous. To see this, let  $\bar{a}_n \rightarrow \bar{a}$  be a convergent sequence of action profiles and let  $a_{i_n} = \varphi_i(\bar{a}_n)$ . This means that  $U(a_{i_n}, (\bar{a}_n)_{-i}) \geq U(b_i, (\bar{a}_n)_{-i})$  for all  $b_i \in A_i$ . Since  $A_i$  is a compact set,  $a_{i_n}$  has a convergent subsequence. Denoting by  $a_i$  the limit of this subsequence and applying limits to the above inequality, we obtain that

$$U(a_i, \bar{a}_{-i}) \geq U(b_i, \bar{a}_{-i}) \quad \text{for all } b_i \in A_i,$$

namely  $a_i = \varphi_i(\bar{a})$ . Since this is true for every convergent subsequence of  $a_{i_n}$ , we have that  $\varphi_i(\bar{a}_n) = a_{i_n} \rightarrow a_i = \varphi_i(\bar{a})$ , which means that  $\varphi$  is continuous.

Now define  $\varphi: A \rightarrow A$  by  $\varphi = (\varphi_1, \dots, \varphi_N)$ . Clearly,  $\varphi$  is a continuous function mapping a compact set to itself. Therefore, by Brouwer’s fixed point theorem, it has a fixed point:  $\varphi(\bar{a}) = \bar{a}$ . We now show that  $\bar{a}$  is a Nash equilibrium of the game. Assume not. Then, there is some  $i \in N$  with  $a_i \in A_i$  such that  $U_i(a_i, \bar{a}_{-i}) - U_i(\bar{a}) = E > 0$ . Then, by concavity of  $U_i$ , for all  $0 < \epsilon < 1$ ,

$$\begin{aligned} & U_i(\epsilon a_i + (1-\epsilon)\bar{a}_i, \bar{a}_{-i}) - U_i(\bar{a}) \\ &\geq \epsilon U_i(a_i, \bar{a}_{-i}) + (1-\epsilon)U_i(\bar{a}) - U_i(\bar{a}) \\ &\geq \epsilon E > 0, \end{aligned}$$

while  $\|\epsilon a_i + (1-\epsilon)\bar{a}_i - \bar{a}_i\|^2 = \epsilon^2 \|a_i - \bar{a}_i\|^2 < \epsilon E$ , for small enough  $\epsilon$ . Therefore, for such small  $\epsilon$ , the action  $\epsilon a_i + (1-\epsilon)\bar{a}_i$  satisfies

$$U_i(\epsilon a_i + (1-\epsilon)\bar{a}_i, \bar{a}_{-i}) - \|\epsilon a_i + (1-\epsilon)\bar{a}_i - \bar{a}_i\|^2 > U_i(\bar{a})$$

which contradicts the fact that  $\varphi_i(\bar{a}) = \bar{a}_i$ .  $\square$

### Mixed Strategies

So far, we have formally defined a game, and have introduced the solution concept of Nash equilibrium which is arguably the central solution concept of game theory. However, there seem to be two problems with this concept. One is that although Nash equilibria exist in a wide class

of games, there are many simple games that do not have a Nash equilibrium. The most troubling example is Matching Pennies. If game theory cannot provide a prediction for this simple game then one must wonder if there is any value to the theory. The second problem is that the concept of Nash equilibrium predicts a very unrealistic outcome in the war of attrition. One would expect that game theory would not only provide nonempty predictions, but also ones that look reasonable and help explain what we see around us.

One way to approach these problems is not to abandon the theory or the concept of Nash equilibrium altogether, but to modify the way we model the problematic situations. The idea behind mixed strategies is to first modify the game by extending the set of actions available to the players, and then to apply the concept of Nash equilibrium to this extended game. In this way one may obtain additional Nash equilibria, some of which may provide reasonable predictions to the game.

Let  $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  be a finite game. For any  $A_i$ , a probability distribution on  $A_i$  is a function

$$x_i: A_i \rightarrow \mathbb{R}_+$$

such that

$$\sum_{a_i \in A_i} x_i(a_i) = 1.$$

The set of all probability distributions on  $A_i$  is denoted by  $\Delta(A_i)$ . A *mixed strategy* on  $A_i$  is a random choice over elements of  $A_i$ , namely an element of  $\Delta(A_i)$ . If  $x_i$  is a mixed strategy on  $A_i$ ,  $x_i(a_i)$  denotes the probability that action  $a_i \in A_i$  is selected when  $x_i$  is adopted. Since elements of  $\Delta(A_i)$  can have an alternative interpretation, such as beliefs about the choice of player  $i$ , we denote the set of mixed strategies by  $X_i$  to distinguish it from the more abstract set of probability distributions on  $A_i$ . Also, we denote the set of mixed strategy profiles as  $X = \times_{i \in N} X_i$ . Denoting for each player  $i \in N$ ,  $X_{-i} = \times_{k \in N \setminus \{i\}} X_k$ , a typical mixed strategy profile can be written as  $(x_k)_{k \in N} = (x_i, x_{-i}) \in X_i \times X_{-i}$ . The *mixed extension* of the strategic game  $G$  is the strategic game  $\langle N, (X_i)_{i \in N}, (U_i)_{i \in N} \rangle$  where the set of actions of player  $i$  is the set of mixed strategies,  $X_i$ , and the payoff function  $U_i: \times_{i \in N} X_i \rightarrow \mathbb{R}$  of player  $i$  is defined by

$$U_i((x_k)_{k \in N}) = \sum_{a=(a_k)_{k \in N} \in A} u_i(a) \prod_{k \in N} x_k(a_k).$$

**Remark 1** Since each mixed strategy of player  $i$ ,  $x_i$ , can be identified with a vector  $x_i = (x_i(a_i))_{a_i \in A_i} \in \mathbb{R}^{|A_i|}$ , the

function  $U_i$  is multinomial in the coordinates of its variables, and, as a result, it is continuous as a function of the players' mixed strategies.

**Definition 4** An *equilibrium in mixed strategies* of the game  $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  is a Nash equilibrium of the mixed extension of the game. In other words, it is a list of mixed strategies  $(x_k^*)_{k \in N} \in X$  such that for all players  $i \in N$  and for all his mixed strategies  $x_i$ ,

$$U_i((x_k^*)_{k \in N}) \geq U_i((x_i, x_{-i}^*)).$$

Alternatively,  $(x_k^*)_{k \in N} \in X$  is a mixed strategy equilibrium if

$$x_i^* \in \mathcal{B}_i(x_{-i}^*) \quad \text{for all } i \in N.$$

Note that for every finite game  $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ , its mixed extension is a strategic game that satisfies the conditions of Theorem 1. As a result, every finite game has a mixed strategy equilibrium.

**Example 1** Consider again Matching Pennies. Its mixed extension is the game  $\langle N, (X_i)_{i \in N}, (U_i)_{i \in N} \rangle$  where the set of players is  $N = \{1, 2\}$ , the sets of mixed strategies are  $X_1 = \{(p_H, p_T) \geq (0, 0): p_H + p_T = 1\}$ , and  $X_2 = \{(q_H, q_T) \geq (0, 0): q_H + q_T = 1\}$ , and the utility functions are given by  $U_1((p_H, p_T), (q_H, q_T)) = p_H q_H + p_T q_T - p_H q_T - p_T q_H$  and  $U_2((p_H, p_T), (q_H, q_T)) = p_H q_T + p_T q_H - p_H q_H - p_T q_T$ . It can be checked that the only Nash equilibrium of this mixed extension is  $((1/2, 1/2), (1/2, 1/2))$ . Indeed, since  $U_1((p_H, p_T), (1/2, 1/2))$  is identically 0, it attains its maximum at, among other strategies,  $(1/2, 1/2)$ . The same is true for  $U_2((1/2, 1/2), (q_H, q_T))$ . To see that there is no other equilibrium, note that for  $(q_H, q_T)$  with  $q_H > q_T$ , player 1's best response is  $(1, 0)$ . But player 2's best response to  $(1, 0)$ , is  $(0, 1)$ . Since  $0 \leq 1$ ,  $(q_H, q_T)$  with  $q_H > q_T$  cannot be part of an equilibrium. Similarly, for any  $(q_H, q_T)$  with  $q_H < q_T$ , player 1's best response is  $(0, 1)$ . But player 2's best response to  $(0, 1)$  is  $(1, 0)$ . Since  $1 \geq 0$ ,  $(q_H, q_T)$  with  $q_H < q_T$  cannot be part of an equilibrium.

We next present a characterization of the mixed strategy equilibria of a game that will sometimes allow us to compute them in an easy way. Further, this characterization serves as the basis of an interesting interpretation of the mixed strategy equilibrium concept that we will discuss later. For this purpose, we identify the action  $a_i \in A_i$  of player  $i$  with the mixed strategy of player  $i$  that assigns probability 1 to action  $a_i$ , and 0 to all other actions. There-

fore, given a player  $i$ , one of his actions  $a_i \in A_i$ , and a profile  $x = (x_k)_{k \in N}$  of the players' mixed strategies,  $(a_i, x_{-i})$  denotes the mixed strategy profile obtained from  $x$  by replacing  $i$ 's mixed strategy  $x_i$  by the mixed strategy of player  $i$  that assigns probability 1 to action  $a_i$ . With this notation we can state the following identity:

$$U_i((x_k)_{k \in N}) = \sum_{a_i \in A_i} x_i(a_i) U_i((a_i, x_{-i})). \quad (1)$$

Indeed,

$$\begin{aligned} U_i((x_k)_{k \in N}) &= \sum_{a=(a_k)_{k \in N} \in A} u_i(a) \prod_{k \in N} x_k(a_k) \\ &= \sum_{a_i \in A_i} \sum_{a_{-i} \in A_{-i}} u_i(a) \prod_{k \in N} x_k(a_k) \\ &= \sum_{a_i \in A_i} x_i(a_i) \sum_{a_{-i} \in A_{-i}} u_i(a) \prod_{k \in N \setminus \{i\}} x_k(a_k) \\ &= \sum_{a_i \in A_i} x_i(a_i) U_i((a_i, x_{-i})). \end{aligned}$$

Identity (1) is useful to prove the following characterization of the mixed strategy Nash equilibria.

**Lemma 1** *The strategy profile  $x^* = (x_k^*)_{k \in N}$  is an equilibrium of the mixed extension of  $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  if and only if for all players  $i \in N$  and for all  $a_i \in A_i$ ,*

$$\text{If } x_i^*(a_i) > 0 \text{ then } U_i((a_i, x_{-i}^*)) = U_i(x^*) \quad (2)$$

$$\text{If } x_i^*(a_i) = 0 \text{ then } U_i((a_i, x_{-i}^*)) \leq U_i(x^*). \quad (3)$$

*Proof* Assume that  $x^* = (x_k^*)_{k \in N}$  satisfies conditions (2) and (3). Let  $i \in N$ , and let  $x_i$  be a mixed strategy of player  $i$ . Then, by (1)

$$\begin{aligned} U_i(x_i, x_{-i}^*) &= \sum_{a_i \in A_i} x_i(a_i) U_i((a_i, x_{-i}^*)) \\ &\leq \sum_{a_i \in A_i} x_i(a_i) U_i(x^*) = U_i(x^*) \end{aligned}$$

and therefore  $x^*$  is an equilibrium.

Assume now that  $x^* = (x_k^*)_{k \in N}$  is an equilibrium. Let  $i \in N$ . Then

$$U_i(x^*) \geq U_i((a_i, x_{-i}^*)) \quad \forall a_i \in A_i \quad (4)$$

and, in particular, condition (3) holds for all  $a_i \in A_i$  such that  $x_i(a_i) = 0$ . Also, using (1) we can write

$$\sum_{a_i \in A_i} x_i^*(a_i) U_i(x^*) = \sum_{a_i \in A_i} x_i^*(a_i) U_i(a_i, x_{-i}^*). \quad (5)$$

If there is  $a_i \in A_i$  such that  $x_i^*(a_i) > 0$  and  $U_i(x^*) > U_i(a_i, x_{-i}^*)$  then, using (4),

$$\sum_{a_i \in A_i} x_i^*(a_i) U_i(x^*) > \sum_{a_i \in A_i} x_i^*(a_i) U_i(a_i, x_{-i}^*)$$

in contradiction to (5).  $\square$

**Corollary 1** *The strategy profile  $x^* = (x_k^*)_{k \in N}$  is an equilibrium of the mixed extension of  $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  if and only if for all players  $i \in N$  and for all  $a_i \in A_i$ ,*

$$x_i^*(a_i) > 0 \text{ implies } a_i \in \mathcal{B}_i(x_{-i}^*).$$

According to the standard interpretation, a player's mixed strategy in a game  $G$  is an action, but in a different game, namely in the mixed extension of  $G$ . According to this interpretation, a mixed strategy is a deliberate choice of a player to use a random device. A mixed strategy equilibrium then is a profile of *independent* random devices, each of which is a best response to the others. Corollary 1 provides an alternative interpretation of a mixed strategy equilibrium. According to this interpretation, a player's mixed strategy represents the uncertainty in the minds of the other players concerning the player's action. In other words, a player's mixed strategy is interpreted not as a deliberate choice of the player but the belief, shared by all the other players, about the player's choice. That is, if  $(x_k)_{k \in N}$  is a profile of mixed strategies, then  $x_i$  is the conjecture, shared by all the players other than  $i$ , about  $i$ 's ultimate choice of action. Consequently,  $x_{-i}$  are the conjectures entertained by player  $i$  about his opponents' actions. According to this interpretation, Corollary 1 says that a mixed strategy equilibrium  $(x_k^*)_{k \in N}$  is a profile of beliefs about each player's actions (entertained by the other players) according to which each player chooses an action that is a best response to his own beliefs.

### The War of Attrition (cont.)

We have seen in Sect. "Analysis of Some Finite Games" that all the Nash equilibria of the war of attrition predict no real fight for the prey. We will now see that there is a mixed strategy equilibrium of the war of attrition that predicts a positive-length fight with probability one.

The players' action sets in the war of attrition are intervals of real numbers. A mixed strategy for player  $i$  in that game can be represented by a cumulative distribution function  $F_i: [0, \infty] \rightarrow [0, 1]$ . For each  $t \in (0, \infty]$ ,  $F_i(t)$  is the probability that player  $i$  gives up at or before  $t$ . We will look for a Nash equilibrium  $(F_1, F_2)$  that consists of two strictly increasing, differentiable cumulative distribution



functions. The density of  $F_i$  is denoted by  $f_i$ . We will try to find an equilibrium at which each player is indifferent between all pure actions.

Consider player  $i$ . Given that his opponent is using mixed strategy  $F_j$ ,  $j \neq i$ , if he chooses to give in at time  $t$ , then he will face a lottery according to which,

- With probability  $1 - F_j(t)$ , player  $i$  does not obtain the prey and gets a payoff of  $-t$ ,
- With probability  $F_j(t)$ , player  $i$  obtains the prey at time  $t_j$ , where  $t_j$  is a random variable whose cumulative distribution function is  $F_j(t_j)/F_j(t)$  (the distribution player  $j$ 's surrender time, conditional on his having given in before  $t$ ).

Therefore, the corresponding expected utility of choosing time  $t$  is

$$\begin{aligned} U_i(t, F_j) &= (1 - F_j(t))(-t) + F_j(t) \int_0^t (v_i - t_j) d \frac{F_j(t_j)}{F_j(t)} \\ &= (1 - F_j(t))(-t) + \int_0^t (v_i - t_j) dF_j(t_j). \end{aligned}$$

Since in the equilibrium we are looking for, player  $i$  is indifferent among all his actions, the above expression is independent of  $t$ . Namely,  $U_i(t, F_j) \equiv c$ . As a result, the derivative of the above utility with respect to  $t$  equals 0. Formally,

$$\begin{aligned} \frac{\partial U_i(t, F_j)}{\partial t} &= t f_j(t) - (1 - F_j(t)) + (v_i - t) f_j(t) \\ &= (1 - F_j(t)) + v_i f_j(t) = 0. \end{aligned}$$

This is a differential equation whose general solution is

$$F_j(t) = 1 - K e^{-\frac{t}{v_i}}.$$

If we want it to satisfy  $F_j(0) = 0$ , we obtain that  $K = 1$ . As a result, the distribution function is given by

$$F_j(t) = 1 - e^{-\frac{t}{v_i}}.$$

Consequently, the equilibrium we are looking for is

$$(F_1(t), F_2(t)) = \left(1 - e^{-\frac{t}{v_2}}, 1 - e^{-\frac{t}{v_1}}\right).$$

According to this equilibrium, for any  $t$ , the probability that there is a fight that lasts at least  $t$  is  $(1 - F_1(t))(1 - F_2(t)) > 0$ . Consequently, there is a fight with probability one. The introduction of mixed strategies allowed the concept of Nash equilibrium to be consistent with fights that last a positive length of time. However, the mixed strategy equilibrium has the following unfortunate property. If

$v_1 < v_2$ , then for all  $t > 0$ ,  $F_1(t) < F_2(t)$ . In other words, it is more likely that the player with the highest willingness to fight for the prey gives up earlier than any given  $t$ , than that the player with the lowest willingness to fight gives in earlier than the same  $t$ . Therefore, in equilibrium it is more likely that the player with the lower willingness to fight wins the war than the other way around. In particular, the probability that player 1 gets the object is given by

$$\int_0^\infty F_2(t) dF_1(t)$$

which can be checked to be equal to  $\frac{v_2}{v_1 + v_2} > 1/2$ . In order to obtain the more intuitive result that the higher the willingness to fight for the prey, the higher is the probability to obtain it, we will need to model the war of attrition in yet a different way. We'll return to this when we introduce asymmetric information to the games.

### Equilibrium in Beliefs

The mixed extension of the game  $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  is constructed in two steps. First, we enlarge the set of actions available to each player by allowing him to choose any mixed strategy on his original action set. Second, since the action choices are now probability distributions over actions, we extend the players' original preferences to preferences over profiles of mixed strategies. We do so by evaluating each mixed strategy profile according to the expected value of the original utilities with respect to the probability distribution over action profiles induced by the mixed strategy.

The first step seems uncontroversial since it is certainly possible for players to use random devices. But the second step is somewhat problematic because, by evaluating mixed strategies according to the expected utility of the resulting lotteries, one is implicitly imposing on the players a certain kind of risk preferences. One may wonder what the implications would be if instead of extending the preferences by assuming that players are expected utility maximizers, we assume that players have more general preferences over profiles of mixed strategies. In particular, we would like to know if there is a suitable generalization of Corollary 1.

Let  $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  be a finite game. We define the mixed extension of  $G$  as the strategic game  $\langle N, (X_i)_{i \in N}, (U_i)_{i \in N} \rangle$  where, as in Sect. "Mixed Strategies",  $X_i$  is the set of probability distributions over the actions in  $A_i$ , for  $i \in N$ , but unlike there, the utility function  $U_i: X \rightarrow \mathbb{R}^N$  is not necessarily a multilinear function of the probabilities, but a general continuous function of the

mixed strategies. The only requirement on  $U_i$  is that for all profiles of degenerate mixed strategies  $(a_k)_{k \in N}$ , we have  $U_i((a_k)_{k \in N}) = u_i((a_k)_{k \in N})$ . As before, a mixed strategy Nash equilibrium of  $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  is a Nash equilibrium of its mixed extension  $\langle N, (X_i)_{i \in N}, (U_i)_{i \in N} \rangle$ . In other words, it is a list of mixed strategies  $(x_k^*)_{k \in N}$  such that for all players  $i \in N$  and for all of his mixed strategies  $x_i$ ,

$$U_i((x_k^*)_{k \in N}) \geq U_i((x_i, x_{-i}^*)).$$

Alternatively,  $(x_k^*)_{k \in N}$  is a mixed strategy equilibrium if

$$x_i^* \in \mathcal{B}_i(x_{-i}^*) \quad \text{for all } i \in N.$$

**Observation 1** It is important to note that two different actions of a player may be best responses to a given mixed strategy profile of the other players, and yet no probability mixture of the two actions will be a best response to the given mixed strategy profile. This will typically be the case when the function  $U_i$  is strictly convex in  $X_i$ , since strictly convex functions attain their maximum at boundary points.

Theorem 1 shows that Nash equilibria exist when the extended utility function  $U_i$  is concave in  $X_i$ . However, Observation 1 indicates that a Nash equilibrium may fail to exist when  $U_i$  is strictly convex in  $X_i$ . Indeed, take a game  $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  with no pure strategy Nash equilibrium, like Matching Pennies, and consider its mixed extension  $\Gamma = \langle N, (X_i)_{i \in N}, (U_i)_{i \in N} \rangle$  where for all players, their extended utility function is strictly convex. Then, for any player  $i \in N$  and for any profile of mixed strategies  $x_{-i}$  of the other players, the set of  $i$ 's best responses  $\mathcal{B}_i(x_{-i})$  consists of only degenerate mixed strategies. Since  $G$  has no pure strategy Nash equilibrium, we conclude that  $\Gamma$  does not have a Nash equilibrium.

**Observation 2** It is also important to note that, unlike in the standard expected utility case, a player's mixed strategy  $x_i^*$  may very well be a best response to some profile  $x_{-i}^*$  of the other players' mixed strategies and at the same time may assign positive probability to an action that (when regarded as a degenerate mixed strategy) is not a best response to  $x_{-i}^*$ . Formally, it may very well be the case that

$$U_i((x_k^*)_{k \in N}) \geq U_i((x_i, x_{-i}^*)) \quad \text{for all } x_i \in X_i$$

and yet

$$U_i((a_i, x_{-i}^*)) < U_i((x_k^*)_{k \in N})$$

for some  $a_i$  such  $x_i^*(a_i) > 0$ .

This will typically occur when the function  $U_i$  is strictly concave in  $X_i$ .

The definition of mixed strategy equilibrium requires from each strategy in the equilibrium profile that it be a best response to the other strategies. Corollary 1 stated that when preferences have the expected utility form, each mixed strategy in a mixed strategy equilibrium is also a probability mixture over best responses to the other strategies in the profile. This result allowed us to interpret a mixed strategy Nash equilibrium as a profile of beliefs, rather than as a profile of probability mixtures. As explained in Observation 2, however, when preferences over mixed strategies are not expected utility preferences, a mixture over best responses is not necessarily a best response. Therefore, Corollary 1 does not extend to the mixed extension where preferences are not of the expected utility form.

In this setup, however, one can still interpret a player's mixed strategy as a belief entertained by the other players about the actions chosen by that player. And a profile of such beliefs will be in equilibrium if the probability distribution over the player's actions that represents  $i$ 's beliefs is obtained as a mixture of best responses of this player to his beliefs about the other players' actions. With this idea in mind, Crawford [8] defined the notion of an equilibrium in beliefs. Before we formally present his definition we need to introduce some notation.

Since when the extended utility functions  $U_i$  are concave in  $i$ 's own strategy a best response to a given profile of the other players' strategies may be a non-degenerate mixed strategy, a mixture of best responses will typically be a mixture over non-degenerate mixed strategies. This mixture induces a probability distribution over actions in a natural way by reducing the compound mixture to a simple mixture. This induced probability distribution can be interpreted as a belief over the actions ultimately chosen. For example, in Matching Pennies, if player 1 believes that there is a probability of 1/2 that player 2 will choose the mixed strategy (1/3, 2/3) and a probability of 1/2 that player 2 will choose the mixed strategy (2/3, 1/3), then player 1 believes that player 2 will choose each one of his two actions with equal probability. More generally, if player  $i$  assigns probability  $p_k$  to the event that player  $j$  will choose mixed strategy  $x^k \in X_j$ , for  $k = 1, \dots, K$ , then player  $i$ 's beliefs about player  $j$ 's actions are given by  $\sum_{k=1}^K p_k x^k \in X_j$ . That is, for each action  $a_j \in A_j$  of player  $j$ , player  $i$  believes that player  $j$  will choose  $a_j$  with probability  $\sum_{k=1}^K p_k x^k(a_j)$ . For each set  $T \subset X_i$  of mixed strategies, let  $D[T] \subset X_i$  denote the set of probability distributions over  $i$ 's actions that are in-

duced by mixtures over elements of  $T$ . With this notation in hand, we can define the concept of equilibrium in beliefs.

**Definition 5** Let  $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  be a game. For each  $i \in N$ , let  $\mathcal{B}_i: X \rightarrow X_i$  be the best response correspondence in the mixed extension of  $G$ . The profile of beliefs  $(x_k^*)_{k \in N} \in \times_{k \in N} \Delta(A_k)$  is an *equilibrium in beliefs* if

$$x_i^* \in D[\mathcal{B}_i(x_{-i}^*)] \quad \text{for all } i \in N.$$

An equilibrium in beliefs is a profile of beliefs  $(x_k^*)_{k \in N}$ . For each  $i \in N$ ,  $x_i^*$  is the common belief of the players other than  $i$  about player  $i$ 's choice of actions. In order for this profile of beliefs to be in equilibrium, we require that for each player  $i \in N$  all the other players believe that  $i$  chooses a mixed strategy that is a best response to his beliefs, which are given by  $(x_k^*)_{k \in N \setminus \{i\}}$ , about the other players' choices of actions. In other words,  $x_i^*$  must be a convex combination of best responses of  $i$  to  $(x_k^*)_{k \in N \setminus \{i\}}$ .

**Example 2** Consider again the mixed extension of Matching Pennies  $\langle N, (X_i)_{i \in N}, (U_i)_{i \in N} \rangle$  where the set of players is  $N = \{1, 2\}$ , the sets of mixed strategies are  $X_1 = \{(p_H, p_T) \geq (0, 0): p_H + p_T = 1\}$  and  $X_2 = \{(q_H, q_T) \geq (0, 0): q_H + q_T = 1\}$ , and the utility functions are now given by  $U_1((p_H, p_T), (q_H, q_T)) = (p_H q_H)^2 + (p_T q_T)^2 - p_H q_T - p_T q_H$  and  $U_2((p_H, p_T), (q_H, q_T)) = (p_H q_T)^2 + (p_T q_H)^2 - p_H q_H - p_T q_T$ . Since the utility functions are strictly convex in the players's own mixed strategies, the best response to any strategy of the opponent is a pure strategy. In particular, one can verify that

$$\mathcal{B}_1(q_H, q_T) = \begin{cases} (1, 0) & \text{if } q_H > q_T \\ \{(1, 0), (0, 1)\} & \text{if } q_H = q_T \\ (0, 1) & \text{if } q_H < q_T \end{cases}$$

and

$$\mathcal{B}_2(p_H, p_T) = \begin{cases} (0, 1) & \text{if } p_H > p_T \\ \{(1, 0), (0, 1)\} & \text{if } p_H = p_T \\ (1, 0) & \text{if } p_H < p_T. \end{cases}$$

It can also be verified that  $((p_H^*, p_T^*), (q_H^*, q_T^*)) = ((1/2, 1/2), (1/2, 1/2))$  is an equilibrium in beliefs. Indeed, for both  $i = 1, 2$ ,  $(1/2, 1/2) \in X_i$  is a convex combination of  $(1, 0)$  and  $(0, 1)$ , which are both in  $\mathcal{B}_j(1/2, 1/2)$ ,  $j \neq i$ . In this equilibrium,

1. Player 1 believes that player 2 will choose  $(1, 0)$  with probability  $1/2$ , and  $(0, 1)$  with probability  $1/2$ .

2. Therefore player 1 believes that player 2 will ultimately choose  $H$  and  $T$ , each with probability  $1/2$ .
3. Given these beliefs, player 1's only best replies are  $(1, 0)$  and  $(0, 1)$ , and
4. Player 2 believes that player 1 will choose each one with probability  $1/2$ . As a result,
5. Player 2 believes that player 1 will ultimately choose  $H$  and  $T$  each with probability  $1/2$ .
6. Given these beliefs, player 2's only best replies are  $(1, 0)$  and  $(0, 1)$ , and
1. Player 1 believes that player 2 will choose  $(1, 0)$  with probability  $1/2$ , and  $(0, 1)$  with probability  $1/2$ .

The following result is a direct implication of the definition of an equilibrium in beliefs.

**Proposition 2 (Crawford [8])** Let  $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  be a strategic game, and  $\Gamma = \langle N, (X_i)_{i \in N}, (U_i)_{i \in N} \rangle$  be the mixed extension of  $G$ , where  $U_i$  is continuous but not necessarily multilinear.

1. Every mixed strategy Nash equilibrium of  $G$  is an equilibrium in beliefs.
2. If for all  $i \in N$ ,  $U_i$  is quasiconcave in  $X_i$ , then every equilibrium in beliefs is a mixed strategy Nash equilibrium of  $G$ .

*Proof*

1. Since  $\mathcal{B}_i(x_{-i}^*) \subset D[\mathcal{B}_i(x_{-i}^*)]$  for all  $i \in N$ , every Nash equilibrium is an equilibrium in beliefs.
2. When the utility function  $U_i$  is quasiconcave in  $i$ 's mixed strategy, the set of best responses  $\mathcal{B}_i(x_{-i}^*)$  is a convex set. Therefore,  $D[\mathcal{B}_i(x_{-i}^*)] = \mathcal{B}_i(x_{-i}^*)$ , and any equilibrium in beliefs is a Nash equilibrium.  $\square$

Crawford [8] shows that although some games have no Nash equilibrium, every game has an equilibrium in beliefs.

### Correlated Equilibrium

In the mixed extension of a game, players do not choose their actions directly, but rather choose probability distributions over their action sets according to which the actions are ultimately selected. The important feature about these probability distributions is that they represent independent random variables. The realization of one player's random variable does not give any information about the realization of the other players' random variables. There is nothing in the bare notion of equilibrium, however, that requires players' behavior to be independent. The basic

feature of an equilibrium is that each player is best responding to the behavior of others, and that each player is free to choose any action in his action set. But one thing is that players can, if they so wish, change their behavior without the consent of others, and another different thing is to expect players' choices to be independent. Therefore, one could ask what would happen if the random devices players use to ultimately choose their actions were correlated. In that case, knowledge of the realization of one's random device would provide some partial information about the realization of the other players' random devices, and therefore of their choices. In equilibrium, a player should take this information into account. To illustrate this point, consider the game of Chicken.

|          |           | Driver 2  |          |
|----------|-----------|-----------|----------|
|          |           | Slow Down | Speed up |
| Driver 1 | Slow Down | 6, 6      | 2, 7     |
|          | Speed up  | 7, 2      | 0, 0     |

This game has two pure-action Nash equilibria, and one equilibrium in mixed strategies. According to the mixed strategy Nash equilibrium, each player chooses Slow Down with probability  $2/3$  and Speed Up with probability  $1/3$ . This mixed strategy equilibrium can be implemented by the following random device. Consider two random variables  $S_1$  and  $S_2$ , whose joint distribution is given by the following table:

Driver 1 chooses his action as a function of the realization of  $S_1$  and Driver 2 chooses his action as a function of the realization of  $S_2$ . (Neither player is informed of the realization of the other player's random variable.) In particular, Driver 1 chooses Slow Down if  $S_1 = 1$  and Speed Up otherwise. Similarly, Driver 2 chooses Slow Down if  $S_2 = 1$ , and Speed Up otherwise. Note that according to this pattern of behavior, each player chooses to slow down with probability  $2/3$ . But more importantly, since  $S_1$  and  $S_2$  are independent random variables, knowledge of the realization of one random variable does not give any information about the realization of the other one. Therefore, after Driver 1 learns the realization of  $S_1$ , he still believes that Driver 2 will choose Slow Down with probability  $2/3$  and consequently any choice is optimal, in particular the one

described above. Similarly, after Driver 2 learns the realization of  $S_2$ , he still believes that Driver 1 will choose to slow down with probability  $2/3$ , and his planned behavior continues to be optimal.

But what would happen if the joint distribution of  $S_1$  and  $S_2$ , was not as presented in Table 1, but rather as follows?

|       |   | $S_2$ |       |
|-------|---|-------|-------|
|       |   | 1     | 2     |
| $S_1$ | 1 | $1/3$ | $1/3$ |
|       | 2 | $1/3$ | 0     |

To answer this question, assume that both players still choose their actions according to the previous pattern of behavior: Driver 1 chooses Slow Down if  $S_1 = 1$ , and Speed Up otherwise. The same holds for Driver 2. As a result, it is still true that each player chooses Slow Down with probability  $2/3$  and Speed Up with probability  $1/3$ . However, since this time the conditioning random variables  $S_1$  and  $S_2$  are not independent, knowledge of the realization of  $S_1$  affects the beliefs of Driver 1 about the probability with which Driver 2 chooses his actions. In particular, if  $S_1 = 1$ , Driver 1 updates his beliefs and assigns probability  $1/2$  to Driver 2 choosing either action, and consequently, Driver 1's only optimal action is Slow Down, which is precisely the choice dictated by the above pattern of behavior. Similarly, if  $S_1 = 2$ , Driver 1 should update his beliefs and assign probability one that Driver 2 will choose Slow Down. Consequently, Driver 1's best reply is to follow the above pattern of behavior and choose Speed Up. One can see that, given that the players know that the random variables  $S_1$  and  $S_2$  are correlated and they use this information accordingly, there is no incentive for either of them to deviate from the proposed pattern of behavior. Therefore, we can say that this pattern of behavior is an equilibrium. This notion of a correlated equilibrium was introduced in Aumann [2]. Before we give a formal definition we introduce the concept of a correlated strategy profile, which will play a central role not only in this section, but in the next one as well.

**Definition 6** Let  $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  be a game. A correlated strategy profile in  $G$  consists of

- A finite probability space  $(\Omega, \pi)$
- For each player  $i \in N$ , a partition  $\mathcal{P}_i$  of  $\Omega$  into events of positive probability
- For each player  $i \in N$ , a function  $\sigma_i: \Omega \rightarrow A_i$  which is measurable with respect to  $\mathcal{P}_i$ .

A correlated strategy profile is a description of what players do and know while playing the game  $G$ . The collec-

Static Games, Table 1  
A random device

|       |   | $S_2$ |       |
|-------|---|-------|-------|
|       |   | 1     | 2     |
| $S_1$ | 1 | $4/9$ | $2/9$ |
|       | 2 | $2/9$ | $1/9$ |

tion  $\langle(\Omega, \pi), (P_i)_{i \in N}\rangle$  represents the random devices used by the players to ultimately choose their actions. The underlying probability space that governs the players' random devices is  $(\Omega, \pi)$ .  $\Omega$  is the set of states, and for each state  $\omega$ ,  $\pi(\omega)$  is the probability that  $\omega$  occurs. For each  $i \in N$ , the partition  $P_i$  represents player  $i$ 's information. Each element of the partition represents a different realization of the random device used by  $i$  to choose his action. Two states that belong to the same element of the partition  $P_i$  cannot be distinguished by  $i$ , while two states that belong to different partition cells can be distinguished by him. For each player  $i$ ,  $\sigma_i: \Omega \rightarrow A_i$  is the random variable that describes players  $i$ 's choice of action,  $\sigma_i(\omega)$  being the action chosen by him at state  $\omega$ . The measurability of  $\sigma_i$  with respect to  $P_i$  formalizes the requirement that the actions chosen by player  $i$  depend only on his information about the state of the world. Therefore, for any two states that belong to the same element of his partition, the actions chosen by  $i$  at those states must be the same. That is, for any  $\omega, \omega' \in P \in P_i$  we have  $\sigma_i(\omega) = \sigma_i(\omega')$ .

For example, the correlated strategy profile described earlier for the game of chicken can be formalized as  $\langle(\Omega, \pi), (P_i)_{i \in N}, (\sigma_i)_{i \in N}\rangle$ , where  $N = \{I, II\}$ , and

- $\Omega = \{(1, 1), (1, 2), (2, 1)\}$
- $\pi(\omega) = 1/3$  for all  $\omega \in \Omega$
- $P_I = \{\{(1, 1), (1, 2)\}, \{(2, 1)\}\}$  and  $P_{II} = \{\{(1, 1), (2, 1)\}, \{(1, 2)\}\}$
- $\sigma_I(\omega) = \begin{cases} \text{Slow Down} & \text{if } \omega \in \{(1, 1), (1, 2)\} \\ \text{Speed up} & \text{if } \omega \in \{(2, 1)\} \end{cases}$
- $\sigma_{II}(\omega) = \begin{cases} \text{Slow Down} & \text{if } \omega \in \{(1, 1), (2, 1)\} \\ \text{Speed up} & \text{if } \omega \in \{(1, 2)\} \end{cases}$ .

According to this correlated strategy profile, there are three equally likely states, and the players can distinguish only one component of the state, namely the realization of their random variable. The players' actions are described by the functions  $\sigma_I$  and  $\sigma_{II}$  which depend only on the respective player's information.

In what follows we denote by  $\sigma: \Omega \rightarrow A$  the function that associates with each  $\omega \in \Omega$  the action profile induced by the strategies  $\sigma_k$ , for  $k \in N$ . That is,  $\sigma = (\sigma_k)_{k \in N}$ . Also, for any  $i \in N$ ,  $\sigma_{-i} = (\sigma_k)_{k \in N \setminus \{i\}}$  so that  $\sigma = (\sigma_{-i}, \sigma_i)$ . We are interested in correlated strategy profiles in which no player benefits by altering his behavior. These special profiles are introduced in the following definition.

**Definition 7** Let  $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  be a strategic game. A *correlated equilibrium* of  $G$  is a correlated strategy  $\langle(\Omega, \pi), (P_i)_{i \in N}, (\sigma_i)_{i \in N}\rangle$  such that for every  $i \in N$  and every function  $\tau_i: \Omega \rightarrow A_i$  that is measurable

with respect to  $P_i$ ,

$$\sum_{\omega \in \Omega} \pi(\omega) u_i(\sigma_{-i}(\omega), \sigma_i(\omega)) \geq \sum_{\omega \in \Omega} \pi(\omega) u_i(\sigma_{-i}(\omega), \tau_i(\omega)). \quad (6)$$

The value  $v_i = \sum_{\omega \in \Omega} \pi(\omega) u_i(\sigma_{-i}(\omega), \sigma_i(\omega))$  is player  $i$ 's correlated equilibrium payoff.

In a correlated strategy profile each player plans to condition his choice of action on the realization of a random variable, and the players' random variables may be correlated. A correlated strategy profile is a correlated equilibrium if no player can find an alternative way to condition his choice on the same random device, so that his expected utility is increased. Note that the player presumably chooses his strategy (his way to condition his actions on the outcomes of the random device) before he learns the realization of the device. Nonetheless, he evaluates the outcomes generated by the players' strategies by taking into account the precise correlation of the random devices on which outcomes players are conditioning their behavior.

Although strictly speaking mixed strategy Nash equilibria are not correlated equilibria, they do induce a correlated equilibrium distribution over action profiles. In order to state this claim, we need the following definition.

**Definition 8** Let  $\langle(\Omega, \pi), (P_i, \sigma_i)_{i \in N}\rangle$  be a correlated strategy profile for  $G$ . Its induced probability distribution over action profiles is given by the function  $p: A \rightarrow [0, 1]$  defined by

$$\begin{aligned} p(a) &= \pi(\{\omega \in \Omega: \sigma(\omega) = a\}) \\ &= \sum_{\{\omega \in \Omega: \sigma(\omega) = a\}} \pi(\omega) \quad \text{for all } a \in A. \end{aligned}$$

**Proposition 3** Let  $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  be a strategic game, and let  $x = (x_1, \dots, x_n)$  be a mixed strategy Nash equilibrium of  $G$ . Then, there is a correlated equilibrium  $\langle(\Omega, \pi), (P_i)_{i \in N}, (\sigma_i)_{i \in N}\rangle$  whose induced probability distribution over action profiles is the same as  $x$ 's distribution.

*Proof* Let  $\langle(\Omega, \pi), (P_i)_{i \in N}, (\sigma_i)_{i \in N}\rangle$  be defined as follows:

- $\Omega = A$
- $\pi(a) = \prod_{i \in N} x_i(a_i)$
- $P_i(a) = \{b \in A: b_i = a_i\}$
- $\sigma_i(a) = a_i$ .



We claim that  $((\Omega, \pi), (P_i)_{i \in N}, (\sigma_i)_{i \in N})$  is a correlated equilibrium whose probability distribution is the same as  $x$ 's distribution. Let  $i \in N$ . Since  $x$  is a mixed strategy Nash equilibrium, we know by Lemma 1 that for all  $a_i \in A_i$

$$\begin{aligned} \text{if } x_i(a_i) > 0 \quad & \text{then } U_i(x) = U_i(a_i, x_{-i}) \\ \text{if } x_i(a_i) = 0 \quad & \text{then } U_i(x) \geq U_i(a_i, x_{-i}). \end{aligned}$$

Consequently, for all  $a_i \in A_i$

$$x_i(a_i)U_i(a_i, x_{-i}) \geq x_i(a_i)U_i(b_i, x_{-i}) \quad \text{for all } b_i \in A_i. \quad (7)$$

Now let  $\tau_i: A \rightarrow A_i$  be a function that is measurable with respect to  $P_i$ . Let  $a_{-i} \in A_{-i}$  be a fixed profile of actions for players other than  $i$ . Letting  $b_i = \tau_i(a_i, a_{-i})$ , Eq. (7) implies that

$$x_i(a_i)U_i(a_i, x_{-i}) \geq x_i(a_i)U_i(\tau_i(a_i, a_{-i}), x_{-i}) \quad \text{for all } a_i \in A_i.$$

Adding over all  $a_i \in A_i$ ,

$$\begin{aligned} \sum_{a_i \in A_i} x_i(a_i)U_i(a_i, x_{-i}) \\ \geq \sum_{a_i \in A_i} x_i(a_i)U_i(\tau_i(a_i, a_{-i}), x_{-i}). \end{aligned}$$

Taking into account the definition of  $U_i(a_i, x_{-i})$  and  $U_i(\tau_i(a_i, a_{-i}), x_{-i})$ , and using the measurability of  $\tau_i$  with respect to  $P_i$ , we get

$$\begin{aligned} \sum_{a_i \in A_i} x_i(a_i) \sum_{a_{-i} \in A_{-i}} \left( \prod_{j \in N \setminus \{i\}} x_j(a_j) \right) u_i(a_i, a_{-i}) \\ \geq \sum_{a_i \in A_i} x_i(a_i) \sum_{a_{-i} \in A_{-i}} \left( \prod_{j \in N \setminus \{i\}} x_j(a_j) \right) u_i(\tau_i(a_i, a_{-i}), a_{-i}) \\ \sum_{a \in A} \left( \prod_{j \in N} x_j(a_j) \right) u_i(a_i, a_{-i}) \\ \geq \sum_{a \in A} \left( \prod_{j \in N} x_j(a_j) \right) u_i(\tau_i(a), a_{-i}) \\ \sum_{a \in A} \pi(a) u_i(a_i, a_{-i}) \geq \sum_{a \in A} \pi(a) u_i(\tau_i(a), a_{-i}) \\ \sum_{a \in A} \pi(a) u_i(\sigma_i(a), \sigma_{-i}(a)) \geq \sum_{a \in A} \pi(a) u_i(\tau_i(a), \sigma_{-i}(a)). \end{aligned}$$

This shows that  $((\Omega, \pi), (P_i)_{i \in N}, (\sigma_i)_{i \in N})$  is a correlated equilibrium of  $G$ . Its induced probability distribution over action profiles is

$$\begin{aligned} p(a) &= \pi(\{b \in A: \sigma(b) = a\}) \\ &= \pi(\{b \in A: b = a\}) \\ &= \pi(a) \\ &= \prod_{i \in N} x_i(a_i). \end{aligned}$$

□

Although a correlated strategy profile consists of a randomizing device used by the players, it turns out that the only feature of the device that determines whether or not the correlated strategy profile constitutes a correlated equilibrium is its induced probability distribution over the action profiles. This is shown by the next proposition.

**Proposition 4** *Let  $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  be a finite strategic game. Every correlated equilibrium probability distribution over action profiles can be obtained in a correlated equilibrium of  $G$  in which*

- $\Omega = A$
- $P_i(a) = \{b \in A: b_i = a_i\}$ .

*Proof* Let  $((\Omega', \pi'), (P'_i, \sigma'_i)_{i \in N})$  be a correlated equilibrium of  $G$ . Consider the correlated strategy profile  $((\Omega, \pi), (P_i, \sigma_i)_{i \in N})$  defined by

- $\Omega = A$
- $\pi(a) = \pi'(\{\omega \in \Omega': \sigma'(\omega) = a\})$  for each  $a \in A$
- $P_i(a) = \{b \in A: b_i = a_i\}$  for each  $i \in N$  and for each  $a \in A$
- $\sigma_i(a) = a_i$  for each  $i \in N$ .

It is clear that this correlated strategy profile induces the required distribution over action profiles. Indeed,

$$\begin{aligned} p(a) &= \pi(\{\omega \in \Omega: \sigma(\omega) = a\}) \\ &= \pi(\{a' \in A: a' = a\}) \\ &= \pi(a) \\ &= \pi'(\{\omega \in \Omega': \sigma'(\omega) = a\}). \end{aligned}$$

It remains to show that this profile is a correlated equilibrium. Take a function  $\tau_i: A \rightarrow A_i$  that is measurable with respect to  $P_i$ . Define  $\tau'_i: \Omega' \rightarrow A_i$  by  $\tau'_i(\omega) = \tau_i(\sigma'(\omega)) = \tau_i(\sigma'_{-i}(\omega), \sigma'_i(\omega))$ . The function  $\tau'_i$  is measurable with respect to  $P'_i$ . Indeed, if  $\omega' \in P'_i(\omega)$ , then  $\sigma'_i(\omega') = \sigma'_i(\omega)$  by measurability of  $\sigma'_i$  with respect to  $P'_i$ . Therefore, by definition of  $P_i$ ,  $P_i(\sigma'(\omega')) = P_i(\sigma'(\omega))$ , and both  $\sigma'_i(\omega')$  and  $\sigma'_i(\omega)$  belong to the same element of

$\mathcal{P}_i$ . Since  $\tau_i$  is measurable with respect to  $\mathcal{P}_i$ , we conclude that  $\tau'_i(\omega') = \tau_i(\sigma'_i(\omega')) = \tau_i(\sigma'_i(\omega)) = \tau'_i(\omega)$ .

Also,

$$\begin{aligned} & \sum_{\omega \in \Omega} \pi(\omega) u_i(\sigma_{-i}(\omega), \tau_i(\omega)) \\ &= \sum_{a \in A} \pi(a) u_i(a_{-i}, \tau_i(a)) \\ &= \sum_{a \in A} \sum_{\{\omega \in \Omega' : \sigma'(\omega) = a\}} \pi'(\omega) u_i(\sigma'_{-i}(\omega), \tau_i(\sigma'(\omega))) \\ &= \sum_{a \in A} \sum_{\{\omega \in \Omega' : \sigma'(\omega) = a\}} \pi'(\omega) u_i(\sigma'_{-i}(\omega), \tau'_i(\omega)) \\ &= \sum_{\omega \in \Omega'} \pi'(\omega) u_i(\sigma'_{-i}(\omega), \tau'_i(\omega)). \end{aligned}$$

In particular, for  $\tau_i = \sigma_i$ ,

$$\begin{aligned} & \sum_{\omega \in \Omega} \pi(\omega) u_i(\sigma_{-i}(\omega), \sigma_i(\omega)) \\ &= \sum_{\omega \in \Omega'} \pi'(\omega) u_i(\sigma'_{-i}(\omega), \sigma'_i(\omega)). \end{aligned}$$

Since  $\langle (\Omega', \pi'), (\mathcal{P}'_i, \sigma'_i)_{i \in N} \rangle$  is a correlated equilibrium,

$$\begin{aligned} & \sum_{\omega \in \Omega'} \pi'(\omega) u_i(\sigma'_{-i}(\omega), \sigma'_i(\omega)) \\ & \geq \sum_{\omega \in \Omega'} \pi'(\omega) u_i(\sigma'_{-i}(\omega), \tau'_i(\omega)) \end{aligned}$$

and therefore

$$\begin{aligned} & \sum_{\omega \in \Omega} \pi(\omega) u_i(\sigma_{-i}(\omega), \sigma_i(\omega)) \\ & \geq \sum_{\omega \in \Omega} \pi(\omega) u_i(\sigma_{-i}(\omega), \tau_i(\omega)). \end{aligned}$$

□

### Rationality, Correlated Equilibrium and Equilibrium in Beliefs

As mentioned earlier, Nash equilibrium and correlated equilibrium are two examples of what is known as solution concepts. Solution concepts assign to each game a pattern of behavior for the players in the game. The interpretation of these patterns of behavior is not always explicit, but it is fair to say that they are usually interpreted either as descriptions of what rational people do, or as prescriptions of what rational people should do. There is a growing literature that tries to connect various game theoretic solution

concepts to the idea of rationality. Rationality is generally understood as the characteristic of a player who chooses an action that maximizes his preferences, given his information about the environment in which he acts. Part of the information a player has is represented by his beliefs about the behavior of other players, their beliefs about the behavior of other players, and so on. So when one speaks of the rationality of players, one needs to take into account their *epistemic state*. There is a formal framework which is appropriate for discussing the actions, knowledge, beliefs and rationality of players. Namely, the framework of a correlated strategy profile. As defined in Sect. “[Correlated Equilibrium](#)”, a correlated strategy profile in a game  $G$  consists of

- A finite probability space  $(\Omega, \pi)$
- For each player  $i \in N$  a partition  $\mathcal{P}_i$  of  $\Omega$  into events of positive probability
- For each player  $i \in N$  a function  $\sigma_i: \Omega \rightarrow A_i$  which is measurable with respect to  $\mathcal{P}_i$ .

For the present discussion we interpret a correlated strategy profile  $\langle (\Omega, \pi), (\mathcal{P}_i)_{i \in N}, (\sigma_i)_{i \in N} \rangle$  as a description of the players' behavior and beliefs, as observed by an outside observer. The set  $\Omega$  is the set of possible states of the world and  $\pi$  is the prior probability on  $\Omega$  shared by all the players. For each player  $i \in N$ ,  $\mathcal{P}_i$  is a partition of  $\Omega$  that represents  $i$ 's information. At state  $\omega \in \Omega$ , player  $i$  is informed not of the state that actually occurred, but of the element  $\mathcal{P}_i(\omega)$  of his partition that contains  $\omega$ . Player  $i$  then uses this information and his prior  $\pi$  to update his beliefs about the true state of the world. Finally, the function  $\sigma_i$  represents the actions taken by player  $i$  at each state. In particular,  $\sigma_i(\omega)$  is the action chosen by  $i$  at state  $\omega$ . Although a correlated equilibrium can be interpreted as a correlated strategy profile *prescribed* by a given solution concept (that of a correlated equilibrium), here we want to interpret a correlated strategy profile as a description of what players actually do and believe. Although players cannot freely choose their beliefs (in the same way as they cannot choose their preferences), they can choose their actions. Furthermore, they have no obligation to behave according to the specified correlated strategy profile. However, ultimately players do behave in a certain way and that behavior is what is represented by the given correlated strategy profile.

Once we fix a correlated strategy profile we can address the rationality of the players. Formally,

**Definition 9** Player  $i \in N$  is *Bayes rational* at  $\omega \in \Omega$  if his expected payoff at  $\omega$ ,  $E(u_i(\sigma) | \mathcal{P}_i)(\omega)$ , is at least as large

as the amount  $E(u_i(\sigma_{-i}, a_i) | \mathcal{P}_i)(\omega)$  that he would have got had he chosen action  $a_i \in A_i$  instead of  $\sigma_i(\omega)$ .

In other words, player  $i$  is rational at a given state of the world if the action  $\sigma_i(\omega)$  he chooses at that state maximizes his expected utility given his information,  $\mathcal{P}_i(\omega)$ , and, in particular, given his beliefs about the actions of the other players.

As before, for any finite set  $T$ , let  $\Delta(T)$  be the set of all probability distributions on  $T$ . The beliefs of player  $i$  about the actions of the other players are represented by his conjectures. A conjecture of  $i$  is a probability distribution  $\psi_i \in \Delta(A_{-i})$  over the elements of  $A_{-i}$ . For any  $j \neq i$ , the marginal of  $\psi_i$  on  $A_j$  is the conjecture of  $i$  about  $j$  induced by  $\psi_i$ . Given a correlated strategy profile  $\langle (\Omega, \pi), (P_i)_{i \in N}, (\sigma_i)_{i \in N} \rangle$ , one can determine the conjectures that each player is entertaining at each state of the world about the actions of the other players. These conjectures are given by the following definition.

**Definition 10** Given a correlated strategy profile  $\langle (\Omega, \pi), (P_i, \sigma_i)_{i \in N} \rangle$ , the conjectures of  $i \in N$  about the other players' actions are given by the function  $\phi_i: \Omega \rightarrow \Delta(A_{-i})$  defined by

$$\phi_i(\omega)(a_{-i}) = \frac{\pi[\{\omega' \in \mathcal{P}_i(\omega): \sigma_{-i}(\omega') = a_{-i}\}]}{\pi[\mathcal{P}_i(\omega)]}.$$

For each  $\omega$ ,  $\phi_i(\omega) \in \Delta(A_{-i})$  is the conjecture of  $i$  at  $\omega$ . For  $j \neq i$ , the marginal of  $\phi_i(\omega)$  on  $A_j$  is the conjecture of  $i$  at  $\omega$  about  $j$ 's actions.

Given a correlated strategy profile, we can speak about what each player knows. The object of knowledge are called *events*, which are the subsets of the set of states of the world  $\Omega$ . We say that player  $i$  knows event  $E \subset \Omega$  at state  $\omega$ , if  $\mathcal{P}_i(\omega) \subset E$ . That is,  $i$  knows  $E$  at  $\omega$  if whatever state he deems possible at  $\omega$  is in  $E$ .

The next result, proved by Aumann and Brandenburger [5], shows a remarkable relationship between the rationality of players and the concept of equilibrium in beliefs.

**Theorem 2** Fix a two-person game,  $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ , and let  $\langle (\Omega, \pi), (P_i)_{i \in N}, (\sigma_i)_{i \in N} \rangle$  be a correlated strategy profile for  $G$ . Let  $\psi_1 \in \Delta(A_1)$  and  $\psi_2 \in \Delta(A_2)$  be two conjectures, one about player 1's actions and the other about player 2's actions. Assume that at some state  $\omega \in \Omega$  each player knows that the other is rational and that their conjectures at  $\omega$  are  $(\phi_1(\omega), \phi_2(\omega)) = (\psi_2, \psi_1)$ . Then,  $(\psi_1, \psi_2)$  is an equilibrium in beliefs.

*Proof* The fact that player  $i$  knows at  $\omega$  that  $j$ 's conjecture is  $\psi_j$  means that

$$\mathcal{P}_i(\omega) \subset \{\omega' \in \Omega: \phi_j(\omega')(a_i) = \psi_j(a_i) \text{ for all } a_i \in A_i\}.$$

Therefore

$$\phi_j(\omega)(a_i) = \psi_j(a_i) \text{ for all } a_i \in A_i. \quad (8)$$

Given Proposition 2 and Corollary 1, we need to show that if  $\psi_i(a_i^*) > 0$ ,  $a_i^*$  is a best response to  $\psi_j$ , for  $i, j = 1, 2$ ,  $i \neq j$ . For this purpose, assume that  $\psi_i(a_i^*) > 0$  for some  $a_i^* \in A_i$ . Then, by definition of  $\phi_j$  and (8),  $\phi_j(\omega)(a_i^*) = \pi[\{\omega' \in \mathcal{P}_j(\omega): \sigma_i(\omega') = a_i^*\}] > 0$ . Consequently, there is  $\omega' \in \mathcal{P}_j(\omega)$  such that  $\sigma_i(\omega') = a_i^*$ . Since player  $j$  knows at  $\omega$  that player  $i$  is rational,

$$\begin{aligned} \omega' \in \mathcal{P}_j(\omega) \subset \{\omega'' \in \Omega: E[u_i(\sigma) | \mathcal{P}_i](\omega'') \\ \geq E[u_i(\sigma_{-i}, a_i) | \mathcal{P}_i](\omega'') \text{ for all } a_i \in A_i\}. \end{aligned}$$

Therefore,

$$E[u_i(\sigma) | \mathcal{P}_i](\omega') \geq E[u_i(\sigma_{-i}, a_i) | \mathcal{P}_i](\omega') \text{ for all } a_i \in A_i$$

and since  $\sigma_i: \Omega \rightarrow A_i$  is measurable with respect to  $\mathcal{P}_i$ ,  $\sigma_i(\omega') = a_i^*$  is the action that player  $i$  chooses at all states in  $\mathcal{P}_i(\omega')$ . Then we can write

$$E[u_i(\sigma_{-i}, a_i^*) | \mathcal{P}_i](\omega') \geq E[u_i(\sigma_{-i}, a_i) | \mathcal{P}_i](\omega') \text{ for all } a_i \in A_i.$$

That is, for all  $a_i \in A_i$

$$\begin{aligned} & \sum_{\omega'' \in \mathcal{P}_i(\omega')} \frac{\pi(\omega'')}{\pi(\mathcal{P}_i(\omega'))} u_i(\sigma_{-i}(\omega''), a_i^*) \\ & \geq \sum_{\omega'' \in \mathcal{P}_i(\omega')} \frac{\pi(\omega'')}{\pi(\mathcal{P}_i(\omega'))} u_i(\sigma_{-i}(\omega''), a_i) \\ & \sum_{a_j \in A_j} \sum_{\substack{\omega'' \in \mathcal{P}_i(\omega') \\ \sigma_j(\omega'') = a_j}} \frac{\pi(\omega'')}{\pi(\mathcal{P}_i(\omega'))} u_i(a_j, a_i^*) \\ & \geq \sum_{a_j \in A_j} \sum_{\substack{\omega'' \in \mathcal{P}_i(\omega') \\ \sigma_j(\omega'') = a_j}} \frac{\pi(\omega'')}{\pi(\mathcal{P}_i(\omega'))} u_i(a_j, a_i) \end{aligned}$$

$$\begin{aligned}
& \sum_{a_j \in A_j} \frac{\pi[\{\omega'' \in P_i(\omega') : \sigma_j(\omega'') = a_j\}]}{\pi(P_i(\omega'))} u_i(a_j, a_i^*) \\
& \geq \sum_{a_j \in A_j} \frac{\pi[\{\omega'' \in P_i(\omega') : \sigma_j(\omega'') = a_j\}]}{\pi(P_i(\omega'))} u_i(a_j, a_i) \\
& \sum_{a_j \in A_j} \phi_i(\omega')(a_j) u_i(a_j, a_i^*) \\
& \geq \sum_{a_j \in A_j} \phi_i(\omega')(a_j) u_i(a_j, a_i). \quad (9)
\end{aligned}$$

Since  $\omega' \in P_j(\omega)$  and player  $j$  knows at  $\omega$  that  $i$ 's conjecture is  $\psi_j$ , then

$$\begin{aligned}
\omega' \in P_j(\omega) \subset \{\omega'' \in \Omega : \phi_i(\omega'')(a_j) = \psi_j(a_j) \\
\text{for all } a_j \in A_j\}.
\end{aligned}$$

Therefore  $\phi_i(\omega')(a_j) = \psi_j(a_j)$  for all  $a_j \in A_j$ , or  $\phi_i(\omega') = \psi_j$ . That is,  $i$ 's conjecture at  $\omega'$  about  $j$ 's actions is  $\psi_j$ . Consequently, substituting in (9),

$$\begin{aligned}
\sum_{a_j \in A_j} \psi_j(a_j) u_i(a_j, a_i^*) & \geq \sum_{a_j \in A_j} \psi_j(a_j) u_i(a_j, a_i) \\
& \forall a_i \in A_i.
\end{aligned}$$

That is,  $a_i^*$  is a best response to player  $i$ 's beliefs about  $j$ 's actions.  $\square$

The only assumptions required by Theorem 2 is that players know they are rational, and that they know each other's conjectures. In a correlated strategy profile for a two-player game, there is only one player entertaining a conjecture about the actions of player 1, namely, player 2. Similarly, player 1 is the only one who entertains a conjecture about the actions of player 2. In an  $n$ -person game, with  $n > 2$ , for each player, there is more than one player entertaining a conjecture about his actions. Therefore, since an equilibrium in beliefs consists of a profile of beliefs, each of which is shared by  $n - 1$  players, a generalization of Theorem 2 would require the players' beliefs about player  $i$ 's actions, for  $i \in N$ , to be identical. In order to obtain these common beliefs it is not sufficient to assume that players know each other's conjectures. One need to strengthen this assumption. Also, in an equilibrium in beliefs, the common belief about player  $i$ 's actions assigns positive probability only to best responses to  $i$ 's conjectures about the choices of the other players. Furthermore,  $i$ 's conjectures about the other players' choices is the product of his beliefs about each of the other players. In other words, an equilibrium in beliefs implicitly assumes that players believe that the other players' choices are independent. Aumann and Brandenburger [5] show that one way

to obtain common conjectures and, simultaneously, that players believe that the other players act independently, is to assume that players' conjectures are commonly known. This surprising and deep result is stated in the next theorem.

**Theorem 3** Let  $G = \langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$  be a strategic game, and let  $((\Omega, \pi), (P_i)_{i \in N}, (\sigma_i)_{i \in N})$  be a correlated strategy profile for  $G$ . Also let  $(\psi_i)_{i \in N} \in \times_{i \in N} \Delta(A_{-i})$  be a profile of conjectures, one for each player. Assume that at some state  $\omega \in \Omega$  each player knows that the others are rational. Further, assume that at  $\omega$  their conjectures are commonly known to be  $(\psi_i)_{i \in N}$ . Then, for each  $j$ , all the conjectures  $\psi_i$  of players  $i$  other than  $j$ , induce the same belief  $\varphi_j \in \Delta(A_j)$  about  $j$ 's actions, and the resulting profile of beliefs,  $(\varphi_i)_{i \in N}$ , is an equilibrium in beliefs.

### Rationality and Correlated Equilibrium

The previous result shows a surprising relationship between the players' rationality and the concept of equilibrium in beliefs. If at some state of the world players know that everybody is rational, and if their conjectures are commonly known at that state, then their beliefs about each player's actions are in equilibrium. It is not that their actions constitute an equilibrium, but that their beliefs do. The question that naturally arises is: are there any epistemic conditions on the players that would induce them to play according to equilibrium? To answer this we turn to Aumann [3], where it is stated that if players are rational at every state, then their behavior constitutes a correlated equilibrium. Therefore, in order to obtain an equilibrium behavior, a sufficient condition is not that players be rational, or that they know that they are rational at some particular state, but that their rationality be common knowledge. And if it is common knowledge that all players are rational, then their behavior is not necessarily a Nash equilibrium, but a correlated equilibrium.

**Theorem 4** Let  $G$  be a strategic game, and let  $((\Omega, \pi), (P_i)_{i \in N}, (\sigma_i)_{i \in N})$  be a correlated strategy profile for  $G$ . If each player is rational at each state of the world, then  $((\Omega, \pi), (P_i)_{i \in N}, (\sigma_i)_{i \in N})$  is a correlated equilibrium.

*Proof* Let  $\tau_i: \Omega \rightarrow A_i$  be a function that is measurable with respect to  $P_i$ . Since  $i$  is Bayes rational at  $\omega$

$$E(u_i(\sigma) | P_i)(\omega) \geq E(u_i(\sigma_{-i}, a_i) | P_i)(\omega) \quad \forall a_i \in A_i.$$

That is,

$$\begin{aligned} & \sum_{\omega' \in \mathcal{P}_i(\omega)} \frac{\pi(\omega')}{\pi(\mathcal{P}_i(\omega))} u_i(\sigma_{-i}(\omega'), \sigma_i(\omega')) \\ & \geq \sum_{\omega' \in \mathcal{P}_i(\omega)} \frac{\pi(\omega')}{\pi(\mathcal{P}_i(\omega))} u_i(\sigma_{-i}(\omega'), a_i) \quad \forall a_i \in A_i. \end{aligned}$$

In particular, for  $a_i = \tau(\omega) = \tau(\omega')$  for all  $\omega' \in \mathcal{P}_i(\omega)$ ,

$$\begin{aligned} & \sum_{\omega' \in \mathcal{P}_i(\omega)} \frac{\pi(\omega')}{\pi(\mathcal{P}_i(\omega))} u_i(\sigma_{-i}(\omega'), \sigma_i(\omega')) \\ & \geq \sum_{\omega' \in \mathcal{P}_i(\omega)} \frac{\pi(\omega')}{\pi(\mathcal{P}_i(\omega))} u_i(\sigma_{-i}(\omega'), \tau(\omega')). \end{aligned}$$

Multiplying both sides by  $\pi(\omega)$  and adding over all the elements of  $\mathcal{P}_i$  we get

$$\begin{aligned} & \sum_{\omega \in \Omega} \pi(\omega) u_i(\sigma_{-i}(\omega), \sigma_i(\omega)) \\ & \geq \sum_{\omega \in \Omega} \pi(\omega) u_i(\sigma_{-i}(\omega), \tau(\omega)). \end{aligned}$$

□

## Bayesian Games

Thus far, we have considered static games, which are objects of the form  $\langle N, (A_i)_{i \in N}, (u_i)_{i \in N} \rangle$ . Although these games have many applications, they are not readily suitable for the analysis of situations involving asymmetric information. Indeed, an implicit assumption behind the definition of a static game is that all players have the same information about the relevant aspects of the situation. In particular, all players have the same information about the sets of actions and preferences of all players. A static game seems suitable to model strategic interactions like the prisoner's dilemma, rock scissors paper, and even chess. At the time they choose their actions, all the players have exactly the same information. There might be what is called strategic uncertainty, namely, uncertainty about what the players will do, but there is no uncertainty about the rules of the game and about the preferences of the players. But how would one translate a game of cards like bridge or poker into a static game? In a game of cards, at the time of choosing his actions, each player knows the cards he holds in his hand, but does not know the cards of his opponents. He only has a belief about the cards held by his opponents. In order to make a sound choice, a player will try to predict the actions of his opponents, but for this it is crucial to use his beliefs about the cards they hold. For the same reason, his opponents should use their beliefs about their own

opponents' cards in order to make a sound choice. Thus, the beliefs about the cards held by each player should be part of a description of a game with asymmetric information. Further, in order to predict his opponents' actions, a player also needs to assess his opponents' beliefs about his own cards. This seems to induce an intractable infinite regress of beliefs, and beliefs about beliefs. Harsanyi [12] provided the basic structure to describe and analyze strategic situations where players are asymmetrically informed. This structure is called a Bayesian game.

**Definition 11** A *Bayesian Game* is a system  $\langle N, (\Omega, \mu), (A_i, \mathcal{P}_i, u_i)_{i \in N} \rangle$  where

- $N$  is the set of players
- $\Omega$  is the set of states of nature
- $\mu$  is the players' common prior belief (a probability measure over the set of states)
- $A_i$  is player  $i$ 's set of actions
- $\mathcal{P}_i$  is player  $i$ 's information partition (a partition of  $\Omega$  into sets of positive measure). Each element of the partition is referred to as a player's type.
- $u_i: \times_{i \in N} A_i \times \Omega$  is player  $i$ 's Bernoulli utility function (a function over pairs  $(a, \omega)$  where  $a \in A$  and  $\omega \in \Omega$ , the expected value of which represents the player's preferences among lotteries over the set of such pairs).

The interpretation of a Bayesian game is as follows. The basic uncertainty is represented by the probability space  $(\Omega, \mu)$  of all states of nature and the prior probability over them. Each state represents a realization of all the parametric uncertainty of the model. For instance, in a game of cards, each state represents each of the possible card deals. The information of player  $i \in N$  is represented by his information partition  $\mathcal{P}_i$ . While states in the same element of the partition cannot be distinguished by the player, he can distinguish between states that belong to different partition cells. In a game of cards, for instance, each partition cell represents a particular set of cards dealt to the player. The probability measure  $\mu$  represents the players' prior belief about the state of nature. This prior belief will be used along with the information obtained by each player to form beliefs about the other players' information. The set of actions of player  $i$  is  $A_i$ . Note that there is no loss of generality in assuming that this set does not depend on the state of nature. One can always add unavailable actions and assign them intolerable disutility. Finally,  $u_i$  is the payoff function that associates to each state of nature and action profile a utility level. Note that since the state of the world is unknown to the player at the time of making his choice, a player faces a lottery for any given action profile. The assumption is that the player evaluates this lottery



according to the expected value of  $u_i$  with respect to that lottery.

Let  $\langle N, (\Omega, \mu), (A_i, \mathcal{P}_i, u_i)_{i \in N} \rangle$  be a Bayesian game. A strategy for player  $i \in N$  is a function  $\sigma_i: \Omega \rightarrow A_i$  that is measurable with respect to  $\mathcal{P}_i$ . We denote the set of strategies for player  $i$  by  $\mathcal{B}_i$ . That is,  $\mathcal{B}_i = \{\sigma_i: \Omega \rightarrow A_i: \sigma_i \text{ is measurable w.r.t. } \mathcal{P}_i\}$ . The interpretation of a strategy in a Bayesian game is the usual one. For each state of nature  $\omega \in \Omega$ ,  $\sigma_i(\omega)$  is the action chosen by player  $i$  at  $\omega$ . The measurability requirement imposes that player  $i$ 's actions depend only on his information. If player  $i$  cannot distinguish between two states of nature, then he must choose the same action at both states. Player  $i$  evaluates a profile  $\sigma: \Omega \rightarrow A$  of strategies according to the expected value of  $u_i$  with respect to  $\mu$ .

In order to define an equilibrium notion for Bayesian games we follow the same idea used for the definition of a mixed strategy equilibrium. Namely, we translate the Bayesian game into a standard game, and then define an equilibrium of the Bayesian game as the Nash equilibrium of the induced game.

**Definition 12** A *Bayesian equilibrium* of a Bayesian game  $\langle N, (\Omega, \mu), (A_i, \mathcal{P}_i, u_i)_{i \in N} \rangle$  is a Nash equilibrium of the strategic game:  $\langle N, (\mathcal{B}_i)_{i \in N}, (U_i)_{i \in N} \rangle$  where for each profile  $\sigma: \Omega \rightarrow A$  of strategies,  $U_i(\sigma) = E_\mu[u_i(\sigma(\omega), \omega)]$  is  $i$ 's expected utility with respect to  $\mu$ .

A Bayesian equilibrium of a Bayesian game is a Nash equilibrium of a properly defined static game. As such, conditions for its existence can be derived from Theorem 1. However, in many situations one is interested in particular kinds of equilibria. Specifically, in the analysis of auctions or of the war of attrition, one is often interested in efficient outcomes. In a single object auction, efficient outcomes are characterized by the fact that in equilibrium the object is allocated to the buyer who values it most. According to many standard auction rules, the object goes to the highest bidder. Therefore, in such auctions, to guarantee an efficient outcome, one would need a monotone equilibrium, namely, one in which bidders bids are higher the higher their valuations for the object are. Athey [1] shows conditions under which a Bayesian equilibrium exists where strategies are non-decreasing. The crucial conditions are that the players' types can be represented by a one-dimensional variable, and that, fixing a nondecreasing strategy for each of a player's opponents, this player's expected payoffs satisfies a single-crossing property. This single-crossing property roughly says that if a high action is preferred to a low action for a given type  $t$ , then the same must be true for all types higher than  $t$ . McAdams [16] extended

Athey's result to the case where types and actions are multidimensional and partially ordered.

### The Asymmetric Information Version of the War of Attrition

We have seen that, when applied to the war of attrition, as modeled by a standard strategic game or by its mixed extension, the notion of Nash equilibrium does not yield a satisfactory prediction.<sup>1</sup> In the former case all the equilibria involve no fight, and in the latter case the equilibrium dictates a more aggressive behavior to the player who values the contested object less. In what follows, we analyze the war of attrition as a Bayesian game. That is, we assume that the players are ex-ante symmetric but they have private information about their value for the contested object.

A Bayesian game that represents the war of attrition is given by  $\langle N, \Omega, (A_i, \mu_i, \mathcal{P}_i, u_i)_{i \in N} \rangle$  where

- $N = \{1, 2\}$
- $\Omega = [0, \infty)^2 = \{(v_1, v_2): 0 \leq v_i < \infty, i = 1, 2\}$
- $A_i = [0, \infty)$  for  $i = 1, 2$
- $\mathcal{P}_i(\hat{v}_1, \hat{v}_2) = \{(v_1, v_2) \in \Omega: v_i = \hat{v}_i\}$  for  $i = 1, 2$
- $\mu((v_1, v_2) \leq (\hat{v}_1, \hat{v}_2)) = F(\hat{v}_1) \times F(\hat{v}_2)$
- $u_i((a_1, a_2), (v_1, v_2)) = \begin{cases} -a_i & \text{if } a_i \leq a_j \\ v_i - a_j & \text{if } a_i > a_j \end{cases}$ .

Here the set of types of player  $i$ , for  $i = 1, 2$ , is represented by the player's willingness to fight,  $v_i$ . The players' willingness to fight are drawn independently from the same distribution  $F$ . A state of the world is, therefore, a realization  $(v_1, v_2)$  of the players' types, and at that state, each player is informed only of his type. Finally, the utility of a player is his valuation for the prey, if he obtains it, net of the time spent fighting for it. We are interested in a symmetric equilibrium in which both players use a symmetric, strictly increasing strategy  $\beta: [0, \infty) \rightarrow [0, \infty)$ , where  $\beta(v)$  is the time at which a player with willingness to fight  $v$  is dictated by the equilibrium to give up. Such an equilibrium would imply that types who value the prey more, are willing to fight more. Further, the probability of observing a fight in equilibrium would not be 0 (in fact, it would be 1.)

It turns out that a symmetric equilibrium strategy is given by

$$\beta(v) = \int_0^v \frac{x f(x)}{1 - F(x)} dx,$$

<sup>1</sup>The war of attrition was analyzed in Maynard Smith [14]. For an analysis of the asymmetric information version of the war of attrition, see Krishna and Morgan [13].

where  $f$  denotes the derivative of  $F$ . To see this, assume that player  $j$  behaves according to  $\beta$  and that player  $i$  chooses to give up at  $t$ . Letting  $z$  be the type such  $\beta(z) = t$ , the expected utility of player  $i$  from choosing  $t$  is

$$U(v_i, z) = \int_0^z (v_i - \beta(y))f(y)dy - \beta(z)(1 - F(z)).$$

Taking derivatives with respect to  $z$ , and using the fact that  $\beta'(z) = zf(z)/[1 - F(z)]$  we obtain

$$\begin{aligned} \frac{\partial U}{\partial z}(v_i, z) &= v_i f(z) - \beta'(z)(1 - F(z)) \\ &= (v_i - z)f(z), \end{aligned}$$

which is positive for  $z < v_i$ , and negative for  $z > v_i$ . As a result, the expected utility of player  $i$  with willingness to pay  $v_i$  is maximized at  $z = v_i$ , which implies that the optimal choice is  $\beta(v_i)$ .

Thus, modeling the war of attrition as an asymmetric game has allowed us to find an equilibrium in which players with higher willingness to fight fight more, and there is a non-negligible probability of observing a fight.

### Evolutionary Stable Strategies

The notion of the Nash equilibrium concept involves players choosing actions that maximize their payoffs given the choices of the other players. The usual interpretation of a Nash equilibrium is as a pattern of behavior that rational players *should* adopt. However, Nash equilibria are sometimes interpreted more descriptively as patterns of behavior that rational players *do* adopt. Certainly, rationality of players is neither a necessary condition nor a sufficient one for players to play a Nash equilibrium. The relationship between rationality and the various solution concepts is not apparent and has been the focus of an extensive literature (see, for example, [3,4,5,7]). Nonetheless, the notion of a Nash equilibrium evokes the idea of players consciously making choices with the deliberate objective of maximizing their payoffs. It is therefore quite remarkable that a concept almost identical to that of Nash equilibrium has emerged from the biology literature. This concept describes a population equilibrium where unconscious organisms are programmed to choose actions with no deliberate aim. In this equilibrium, members of the population meet at random over and over again to interact. At each interaction, these players act in a pre-programmed way and the result of their actions is a gain in biological fitness. Fitness is a concept related to the reproductive value or survival capacity of an organism. In a temporary equilibrium, the fitness gains are such that the proportions of

individuals that choose each one of the possible actions remain constant. However, this temporary equilibrium may be disturbed by the appearance of a mutation, which is a new kind of behavior. This mutation may upset the temporary equilibrium if its fitness gains are such that the new behavior spreads over the population. Alternatively, if the fitness gains of the original population outweigh those of the mutation, then the new behavior will fail to propagate and will eventually disappear. In a population equilibrium, the interaction of any mutant with the whole population awards the mutant insufficient fitness gains, and as a result the mutants disappear. The notion of a population equilibrium is formalized by means of the concept of an evolutionary stable strategy, introduced by Maynard Smith and Price [15].

In what follows we restrict our attention to symmetric two-player games. So let  $G = (\{1, 2\}, \{A_1, A_2\}, \{u_1, u_2\})$  be a game such that  $A_1 = A_2 = \bar{A}$ , and such that for all  $a, b \in \bar{A}$ ,  $u_1(a, b) = u_2(b, a)$ . An evolutionary stable strategy is an action in  $\bar{A}$  such that if all members of the population were to choose that action, no sufficiently small proportion of mutants choosing an alternative action would succeed in invading the population. Alternatively, an evolutionary stable strategy is an action in  $\bar{A}$  such that if all the members of the population were to choose that action, the population would reject all sufficiently small mutations involving a different action.

More specifically, suppose that all members of the population are programmed to choose  $a \in \bar{A}$ , and then a proportion  $\varepsilon$  of the population mutates and adopts action  $b \in \bar{A}$ . In that case, the probability that a given member of the population meets a mutant is  $\varepsilon$ , while the probability of meeting a member that plays  $a$  is  $1 - \varepsilon$ . Therefore, the mutation will not propagate and will vanish if the expected payoff of a mutant is less than the expected payoff of a member of the majority. Otherwise it will propagate. This leads to the following definition.

**Definition 13** An action  $a \in \bar{A}$  is an evolutionary stable strategy of  $G$  if there is an  $\bar{\varepsilon} \in (0, 1)$  such that for all  $\varepsilon \in (0, \bar{\varepsilon})$ , and for all  $b \in \bar{A}$

$$(1 - \varepsilon)u_1(a, a) + \varepsilon u_1(a, b) > (1 - \varepsilon)u_1(b, a) + \varepsilon u_1(b, b). \quad (10)$$

The following result shows that the concept of an evolutionary stable strategy is very close to the notion of a Nash equilibrium.

**Proposition 5** If  $a \in \bar{A}$  is an evolutionary stable strategy of  $G$ , then  $(a, a)$  is a Nash equilibrium. And if  $(a, a)$  is

a strict Nash equilibrium then  $a$  is an evolutionary stable strategy.

*Proof* If  $u_1(a, a) > u_1(b, a)$  for all  $b \in \bar{A} \setminus \{b\}$ , then inequality (10) holds for all sufficiently small  $\varepsilon > 0$ . If  $u_1(b, a) > u_1(a, a)$  for some  $b \in \bar{A}$ , the reverse inequality holds for all sufficiently small  $\varepsilon$ .  $\square$

### Future Directions

Static games have been shown to be a useful framework for analyzing and understanding many situations that involve strategic interaction. At present, a large body of literature is available that develops various solution concepts, some of which are refinements of Nash equilibrium and some of which are coarsenings of it. Nonetheless, several areas for future research remain. One is the application of the theory to particular games to better understand the situations they model, for example auctions. In many markets trade is conducted by auctions of one kind or another, including markets for small domestic products as well as some centralized electricity markets where generators and distributors buy and sell electric power on a daily basis. Also, auctions are used to allocate large amounts of valuable spectrum among telecommunication companies. It would be interesting to calculate the equilibria of many real life auctions. Simultaneously, future research should also focus on the design of auctions whose equilibria have certain desirable properties.

Another future direction would be to empirically and experimentally test the theory. The various equilibrium concepts predict certain kinds of behavior in certain games. Our confidence in the predictive and explanatory power of the theory depends on its performance in the field and in the laboratory. Moreover, the experimental and empirical results should provide valuable feedback for further development of the theory. Although some valuable experimental and empirical tests have already been performed (see [17,19,22,25] to name a few), the empirical aspect of game theory in general, and of static games in particular, remains underdeveloped.

### Bibliography

1. Athey S (2001) Single crossing properties and the existence of pure strategy equilibria in games of incomplete information. *Econometrica* 69:861–889
2. Aumann RJ (1974) Subjectivity and correlation in randomized strategies. *J Math Econ* 1:67–96
3. Aumann RJ (1987) Correlated equilibrium as an expression of bayesian rationality. *Econometrica* 55:1–18
4. Aumann RJ (1995) Backward induction and common knowledge of rationality. *Games Econ Behav* 8:6–19
5. Aumann RJ, Brandenburger A (1995) Epistemic conditions for Nash equilibrium. *Econometrica* 63:1161–1180
6. Binmore K (2007) *Playing for Real*. Oxford University Press, New York
7. Brandenburger A, Dekel E (1987) Rationalizability and correlated equilibria. *Econometrica* 55:1391–1402
8. Crawford V (1990) Equilibrium without independence. *J Econ Theory* 50:127–154
9. Fudenberg D, Tirole J (1991) *Game Theory*. MIT Press, Cambridge
10. Geanakoplos J (2003) Nash and Walras equilibrium via Brouwer. *Econ Theory* 21:585–603
11. Glicksberg IL (1952) A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points. *Proc Am Math Soc* 3:170–174
12. Harsanyi J (1967) Games with incomplete information played by 'Bayesian' players. *Manag Sci Part i* 14:159–82; *Part ii* 14:320–34; *Part iii* 14:486–502
13. Krishna V, Morgan J (1997) An analysis of the war of attrition and the all-pay auction. *J Econ Theory* 72:343–362
14. Smith MJ (1974) The theory of games and the evolution of animal conflicts. *J Theor Biol* 47:209–221
15. Smith MJ, Price GR (1973) The logic of animal conflict. *Nature* 246:15–18
16. McAdams D (2003) Isotone equilibrium in games of incomplete information. *Econometrica* 71:1191–1214
17. McKelvey RD, Palfrey TR (1992) An experimental study of the centipede game. *Econometrica* 60:803–836
18. Nash JF (1950) Equilibrium points in  $n$ -person games. *Proc Natl Acad Sci USA* 36:48–49
19. O'Neill B (1987) Nonmetric test of the minimax theory of two-person zero-sum games. *Proc Natl Acad Sci* 84:2106–2109
20. Osborne MJ (2004) *An Introduction to Game Theory*. Oxford University Press, New York
21. Osborne MJ, Rubinstein A (1994) *A Course in Game Theory*. MIT Press, Cambridge
22. Palacios-Huerta I (2003) Professionals play minimax. *Rev Econ Stud* 70:395–415
23. von Neumann J (1928) Zur Theorie der Gesellschaftsspiele. *Math Ann* 100:295–320
24. von Neumann J (1959) On the theory of games of strategy. In: Tucker AW, Luce RD (eds) *Contributions to the Theory of Games*, vol IV. Princeton University Press, Princeton
25. Walker M, Wooders J (2001) Minimax play at Wimbledon. *Am Econ Rev* 91:1521–1538

## Statistical Applications of Wavelets

SOFIA OLHEDE

Department of Statistical Science,  
University College London, London, UK

### Article Outline

Glossary

Definition of the Subject

Introduction

Decomposition View of Random Processes  
Wavelets  
Signal Estimation or Denoising  
2-D Extensions of Wavelets  
Covariance Estimation Using Wavelets  
Future Directions  
Bibliography

## Glossary

**Locally stationary process** A stochastic process that when sampled in time, its sampled points have joint characteristics in a short time window that only depend on the difference between the sampled time points, and not the global time point of the samples within the time window.

**Multiresolution analysis** The analysis of a phenomenon of interest over different temporal or spatial scales and locations.

**Shrinkage** A method of estimation where the initial estimator is decreased by multiplication with a factor of magnitude less than or equal to one. If the object being estimated is small, then a reduced variance will arise, that sets off the increased bias in estimation.

**Sparsity** Sparsity in a vector corresponds to the support of the vector being limited. The sparsity can either be strict, i. e. the vector is perfectly supported in a small subset of all its entries, or the magnitudes of the entries decay sufficiently rapidly, if ordered in magnitude.

**Stationary process** A stochastic process that when sampled in time, its sampled points have joint characteristics that only depend on the difference between the sampled time points, and not the global time point of the samples.

**Thresholding** A method of estimation where the initial estimator is set to zero if it does not exceed a given threshold. Thresholding is a special case of shrinkage.

**Wavelet and scaling functions** The functions that the wavelet and scaling filters converge to in increasing scale, when using discrete wavelet filters. These are not for any finite scale equivalent to the wavelet and scaling filters.

**Wavelet and scaling filters** The high and low pass digital filters that are used to calculate the discrete wavelet transform of a given sequence or vector. These are a pair of quadrature mirror filters that satisfy the perfect reconstruction property.

## Definition of the Subject

A wavelet is a function essentially supported over a small time-window, that is oscillatory, with a local period that

ranges over a limited sets of values. A wavelet function can be considered jointly local in time and frequency. Wavelets have become an integral part of statistical analysis of structured phenomena. Their utility for inference is based on several of their important properties, namely compression, separation of multiscale features, and differencing of a given set of moments. The statistical analysis of processes using wavelets presents a more faceted view of the phenomenon of interest than traditional methods. The utility of this in many areas is unquestionable, and statistical applications of wavelets include areas as disparate as climatology, econometrics, finance, geophysics, mathematics, as well as signal and image processing.

## Introduction

The field of wavelets in statistics can roughly be divided into two areas, namely the usage of continuous wavelets, and the usage of discrete wavelets. Continuous wavelets are defined at a continuum of locations, whilst discrete wavelet filters are sequences defined only at a countable number of locations and scales. These sequences are not strictly speaking wavelet functions, but converge in increasing scale to such functions. Unfortunately in the literature both functions and filters are referred to as ‘wavelets’.

Continuous wavelets are applied to the characterization of random processes, but are not usually used for actual signal estimation, as the transform does not possess an exact discretely implementable inverse. Continuous wavelets do not necessarily satisfy the constraint of compactness in time, and this permits more freedom in their design. A third class of wavelets that often receive separate treatment are complex-valued discrete wavelet filters. These can be thought of as bridging the gap between continuous and discrete representations. Complex wavelets are often chosen to be redundant, and have interpretable magnitude and phase, see for example the review of these in [48].

The statistical representation of structured phenomena using decompositions was first introduced by [47] who analyzed time series using Fourier decompositions. This analysis method corresponds to decomposing a sample into cosines and sines to determine the power associated with each frequency or mode. The interpretation of the observed weight attached to each frequency depends on the model posited for the observed data. If the process is assumed to be *stationary*, then the observed weight can be shown to converge in expectation to the true weight, namely the Fourier transform of the autocovariance of the



time series. By calculating the full set of weights across frequencies, we may obtain a picture of the frequency content of the process. Even for processes that are not *stationary*, decomposing the set of observations into sinusoids allows us to visualize the frequency content of the signal. As long as the process is *harmonizable*, see [36], such decompositions are suitable.

Unfortunately whilst the decomposition of a time-varying process into sines and cosines is interpretable in a global sense, and can be shown to possess some suitable properties, it is not as easily interpreted as might be the case. Many processes are generated by mechanisms that fundamentally alter over time. It is therefore suitable to use decomposition functions that unlike sines are limited in time, but still correspond to oscillatory functions. A popular choice of such functions is wavelets.

A wavelet is a time-limited oscillation, or a “little wave”. The first family was constructed by Alfred Haar in 1910, see [26]. The subject lay dormant for many years, until the mid 1980s when their usage in geophysics revitalized interest in the set of decompositions, by Morlet, and one of the groundbreaking new articles corresponds to [24]. A great theoretical leap forward was the construction of discrete wavelet filters, see [10]. Most of the statistical usage of wavelets has focused around the usage of discrete wavelets.

The development of statistical methods can really be grouped into three main strands of development: i) the usage of compression for non-parametric signal estimation, following work by [15,16,17], ii) the characterization of stochastic difference stationary processes, see [38,51], and iii) the inference of locally stationary wavelet processes, see [40]. Theoretical results have also been achieved in terms of the characterization of stochastic processes, see [3,4,8,34], but these results as of yet have had limited practical implications.

After initial usage of the wavelet transform for signal estimation it became obvious that despite the wavelet transform’s orthogonalizing effects, relationships between coefficients must be used, to derive better representations, see [58]. The development of data driven decompositions, and usage of the ‘lifting algorithm’, [12], followed in the second strand of developments. Further steps forward included the construction of two dimensional decompositions for image estimation, where notable contributions include the ridgelet, curvelet and bandelet transforms, see [9,33,52]. Outstanding problems remain in the statistical usage of these transforms, and investigations have mainly been limited to modeling the observed phenomena as deterministic and immersed in some form of noise.

## Decomposition View of Random Processes

Wavelets are used to analyze observations whose generating mechanism can be characterized in terms of the location the observation is made at. This could be time, space or some other suitable indexing. We shall denote the location of any observation by  $t \in \mathbb{R}^d$ , where  $d$  denotes the dimension of the indexing, this could be  $d = 1$ , like a time series, or  $d = 2$ , like an image. If  $\{x(t)\}$  is a random process then its structure can be determined from its moments. The first order structure of  $x(t)$  is given by its mean, defined by

$$\mu_x(t) = E\{x(t)\}, \quad t \in \mathbb{R}^d, \quad (1)$$

and the second order structure of  $x(t)$  is given by its covariance, namely

$$\gamma_x(t_1, t_2) = \text{cov}\{x(t_1), x(t_2)\}, \quad t_1, t_2 \in \mathbb{R}^d. \quad (2)$$

$E\{\cdot\}$  denotes the expectation operator, and  $\text{cov}\{\cdot, \cdot\}$  the covariance operator. If the process is assumed to be a Gaussian process, i.e. if any sample of  $n$  observations from  $\{x(t)\}$  is multivariate Gaussian, then  $\{\mu_x(t)\}$  in combination with  $\{\gamma_x(t_1, t_2)\}$  fully characterize the generation of the process  $\{x(t)\}$ .

Often we only take  $n$  observations from the process  $\{x(t)\}$  and still want to fully characterize the generating mechanism, so that we can test hypotheses of interest regarding  $\{x(t)\}$  or just estimate some parameters of interest. To be able to do this we must make some further assumptions about  $x(t)$ , and these could correspond to fully parametric or non-parametric assumptions on  $\{\mu_x(t)\}$  or  $\{\gamma_x(t_1, t_2)\}$ . Typical examples might be  $\mu_x(t) = a_x + b_x t$ , which is a parametric specification of the mean, or  $\mu_x(t)$  possessing a certain degree of smoothness. If the latter approach is taken then it is convenient to represent  $\mu_x(t)$  in terms of some functions where the assumed smoothness can be easily incorporated into the modeling. Furthermore the covariance structure can be assumed to simplify if we choose a suitable representation. If  $x(t)$  is represented in terms of linear combinations of known basis functions, such as wavelets, then Gaussian processes will be transformed into Gaussian processes with a different mean and covariance function. For certain classes of auto-covariance functions the analysis is much simplified by considering the transformed rather than the original Gaussian process.

## Wavelets

A wavelet function, is a complex-valued square integrable function satisfying the admissibility condition, see [27]. We denote a generic wavelet by  $\psi(t)$ . The Continuous



Wavelet Transform (CWT) of a signal  $x(t)$  with respect to a wavelet  $\psi(t)$  is defined as a filtration of  $x(t)$ :

$$w^{(x)}(t, s) = \int_{-\infty}^{\infty} x(t') \psi_{t,s}(t') dt',$$

$$\psi_{t,s}(t') = s^{-\frac{1}{2}} \psi\left(\frac{t' - t}{s}\right). \quad (3)$$

The choice of normalization, here  $s^{-1/2}$  varies across application fields. Sometimes  $1/2$  is replaced by  $1$  if more convenient for analysis, see [54]. The former is referred to as the *energy normalization*, and the latter as the *compression normalization*. The former normalization conserves the variance of the transformed process.

If  $\{x(t)\}$  is a stochastic process satisfying some regularity conditions, see [34], then  $\{w^{(x)}(t, s)\}$  is a doubly indexed stochastic process with mean

$$E\{w^{(x)}(t, s)\} = \int_{-\infty}^{\infty} \mu_x(t') \psi_{t,s}(t') dt',$$

$$\psi_{t,s}(t') = s^{-\frac{1}{2}} \psi\left(\frac{t' - t}{s}\right) \quad (4)$$

$$\equiv w^{(\mu_x)}(t, s). \quad (5)$$

The form of this expression simplifies if  $x(t)$  takes the simpler form of a time-varying oscillation, see [13], or a singularity [28,55], or is a sufficiently regular function, see [56]. Under such circumstances very few of  $\{w^{(\mu_x)}(t, s)\}$  are non-zero and inferences can be made about the model from the few non-zero coefficients. In the special case of time-varying oscillations the theory of *wavelet ridge analysis* has been developed to characterize the local oscillations. Applications of such methods include mechanical vibratory signals, see [31,59], seismic signals [44,45], as well as ocean eddy time series, see [6,35].

The usage of continuous wavelets does not extend greatly beyond the aforementioned methods of signal inference. The characterization of various stochastic processes have been investigated using continuous wavelets, i. e. [3,8,25,32,34], but this has been put to little practical usage. Some interesting developments for the prediction of continuous signals using wavelets have been developed in [1,2].

The main focus of wavelet usage in statistics has instead evolved around the usage of the Discrete Wavelet Transform (DWT). The DWT is defined as a reversed convolution between a discretely sampled signal and the wavelet filter. We denote the  $j$ th level wavelet filter  $\{h_{j,l}\}$ , and the scaling filter  $\{g_{j,l}\}$ . For a complete review of the properties of these objects see [11,37,46].

These are constructed from the  $j = 1$  level filters  $\{h_l\}$  and  $\{g_l\}$ , that satisfy the quadrature mirror relationship of  $g_l = (-1)^{l+1} h_{L-1-l}$  where  $L$  is the length of the two filters. Given a sampled process  $x_n = x(n\Delta t)$  the wavelet coefficients are defined by:

$$w_{j,l}^{(x)} = \sum_n h_{j,l} x_n, \quad h_{j,l} = \sum_{k=0}^{L-1} h_k g_{j-1,l-2^{j-1}k}, \quad (6)$$

where the wavelet filters given  $\{g_{j-1,l}\}$  can now be determined iteratively. To complement the wavelet coefficients of  $x_n$ , additionally the scaling coefficients are computed, of

$$v_{j,l}^{(x)} = \sum_n g_{j,l} x_n, \quad g_{j,l} = \sum_{k=0}^{L-1} g_k g_{j-1,l-2^{j-1}k}, \quad (7)$$

this completing the iterative specification of the wavelet and scaling filters. If we collect the wavelet and scaling coefficients in a vector  $\mathbf{W}^{(x)} = [w_{1,0}^{(x)}, \dots, w_{1,N/2-1}^{(x)}, w_{2,0}^{(x)}, \dots, v_{J/2,0}^{(x)}]$ , then the DWT can be represented as:

$$\mathbf{W}^{(x)} = \mathcal{W} \mathbf{X}, \quad (8)$$

where  $\mathcal{W}$  is composed of  $\{h_{j,l}\}$  and  $\{g_{j,l}\}$ , see [46]. Equation (8) is not the method to compute the DWT, popularly the transform is implemented using Mallat's pyramid algorithm, see [37], but it enables the easy determination of the stochastic properties of the wavelet transform. The CWT and the DWT are not two disparate objects: if we possess a sequence of coefficients  $\{v_{0,l}^{(x)}\}$  such that

$$x(t) = \sum_{n=-\infty}^{\infty} v_{0,l}^{(x)} \phi_{0,k}(t) \quad (9)$$

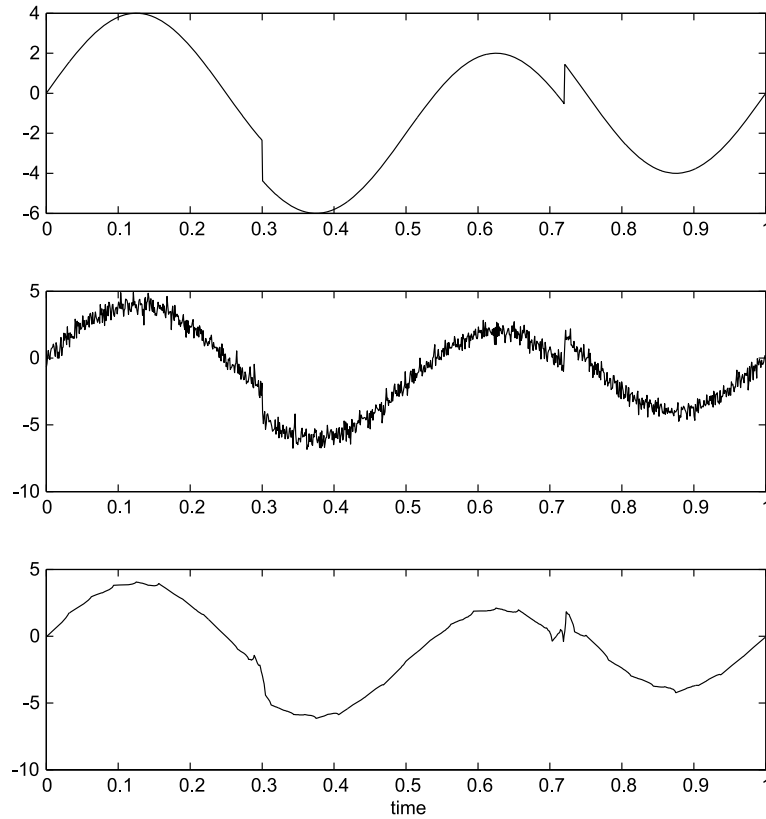
then  $w_{j,l}^{(x)}$  can be calculated from  $v_{0,l}^{(x)}$ , as can  $v_{j,l}^{(x)}$ , and correspond to the CWT as well as scaling coefficients of the continuous rather than the sampled process. In practise  $x_n$  is equated to  $v_{0,n}^{(x)}$ , even if this is formally inappropriate.

### Signal Estimation or Denoising

The most popular statistical application of wavelets is denoising, or function estimation. This proceeds from a model of the observed signal of:

$$x_n = \mu(n\Delta t) + \epsilon_n, \quad n = 0, \dots, N-1, \quad (10)$$

where  $\Delta t$  is referred to as the sampling period of the signal, and the sample size is  $N$ . To fully specify the generation of the observations  $\{\epsilon_n\}$  must be modeled, and a popular model is to assume this sequence as zero-mean, uncorrelated (white) and Gaussian, with some constant vari-



Statistical Applications of Wavelets, Figure 1

The clean Heavisine signal (*top*), the noisy Heavisine signal (*middle*) and the denoised Heavisine signal (*bottom*). The plots show the typical features of a wavelet estimator, namely the adaptation to local smoothness, where the sinusoidal parts of the signal are smoothed, and the discontinuities, are kept without over smoothing

ance  $\sigma^2$ , that needs to be determined to implement estimation. To see a typical realization of such a signal, see Fig. 1. This signal, 'Heavisine' is famous as one of the four typical signals that were introduced by [16]. Heavisine has been immersed in noise, see Fig. 1. This signal is not trivial to smooth as it possesses varying degrees of smoothness across the realization, compare for example the discontinuities with the smooth portions of the signal.

One important inference problem in this setting is to estimate  $\{\mu(n\Delta t)\}$  with low expected square deviation, i. e. to chose an estimator  $\hat{\mu}(n\Delta t)$  such that

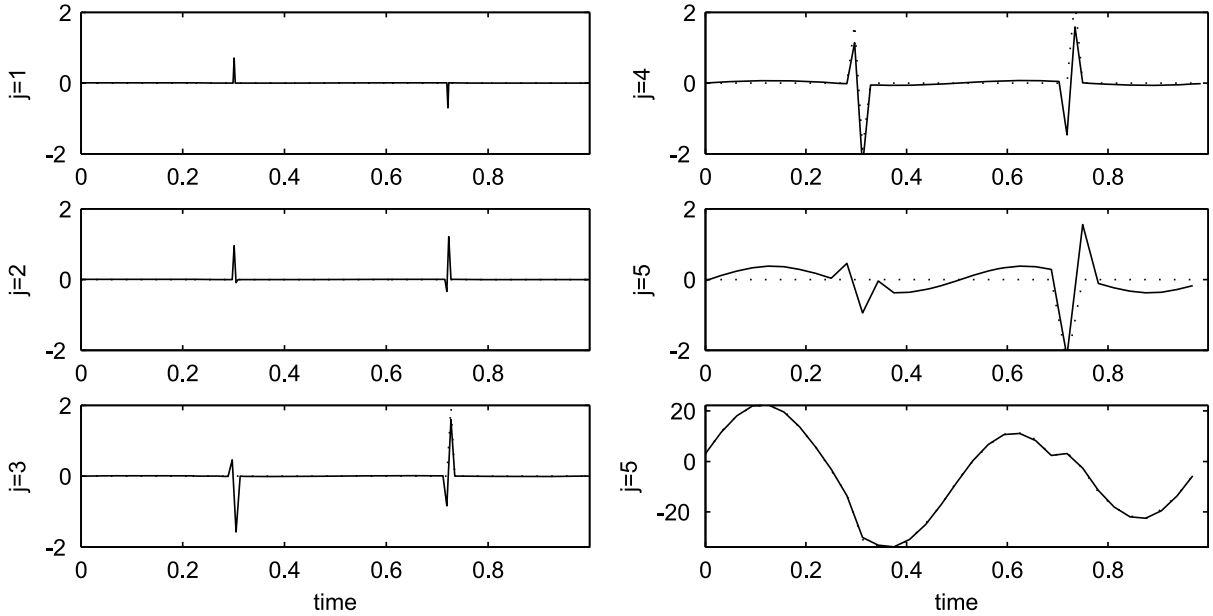
$$\text{MSE}(\hat{\mu}, \mu) = \sum_{n=0}^{N-1} \mathbb{E} \left\{ \left( \mu(n\Delta t) - \hat{\mu}(n\Delta t) \right)^2 \right\}, \quad (11)$$

is as small as possible. Common methods of estimating  $\mu(n\Delta t)$  include using many different adaptive linear estimation methods, such as kernel smoothers, smoothing splines, truncated Fourier methods etc. Smoothing using

the DWT was introduced by [15,16], and has enjoyed an incredible success. The simplest form of wavelet estimation is based on the compression of the DWT of  $x(t)$  alone, combined with the uniformity of the white process across time.

The wavelet coefficients of  $\mu(\cdot)$  are given as  $\{w_{j,l}^{(\mu)}\}$ , and together with the scaling coefficients,  $\{v_{j,l}^{(\mu)}\}$ , these are sufficient to reconstruct  $\mu(n\Delta t)$  via calculating the Inverse DFT (IDFT) or:

$$\begin{pmatrix} \mu(0\Delta t) \\ \vdots \\ \mu((N-1)\Delta t) \end{pmatrix} = \mathcal{W}^{-1} \begin{pmatrix} w_{1,0}^{(\mu)} \\ \vdots \\ w_{1,\frac{N}{2}-1}^{(\mu)} \\ w_{2,0}^{(\mu)} \\ \vdots \\ w_{2,\frac{N}{4}-1}^{(\mu)} \\ \vdots \\ v_{J,0} \end{pmatrix}. \quad (12)$$



Statistical Applications of Wavelets, Figure 2

The wavelet coefficients of the Heavisine signal, used in the original paper by Donoho and Johnstone [15]. We show the first five level wavelet coefficients, and the scaling coefficients of the signal in *solid line*. Added to the picture are the thresholded wavelet and scaling coefficients of a noisy realization in *dotted line*

Thus a good estimator of  $\{w_{j,l}^{(\mu)}\}$  can easily be converted into a good estimator of  $\mu(n\Delta t)$ , using Eq. (12). To see the motivation behind the most commonly adopted choice of estimation for the wavelet coefficients, observe the wavelet coefficients of the Heavisine signal, that we plotted in Fig. 1, plotted in Fig. 3. Most of the wavelet coefficients are quite small. If a given signal  $\mu(\cdot)$  is sufficiently regular then it can be shown that  $\{w_{j,l}^{(\mu)}\}$  must decrease in magnitude. For this reason one may argue that most of this set are zero, and the transform exhibits compression or sparsity, see [56]. We are therefore trying to estimate a sequence of variables of which most are zero, and only a few are large. It is *not* optimal to equate the estimator to the sample DWT, but instead shrinkage estimation is suitable. Consider the generic setting of an estimator for  $w$  denoted by  $\hat{w}$  which is unbiased, but potentially has large variance. A shrinkage estimator of this quantity is given by  $\hat{w}^{(s)} = c\hat{w}$  for some  $0 \leq c \leq 1$ . Clearly this estimator has smaller variance as compared to  $\hat{w}$  (or possibly equal variance), but introduces some bias into the estimation. The mean square error is in fact given by:

$$\text{MSE}(\hat{w}, w) = c^2\sigma^2 + (1-c)^2w^2. \quad (13)$$

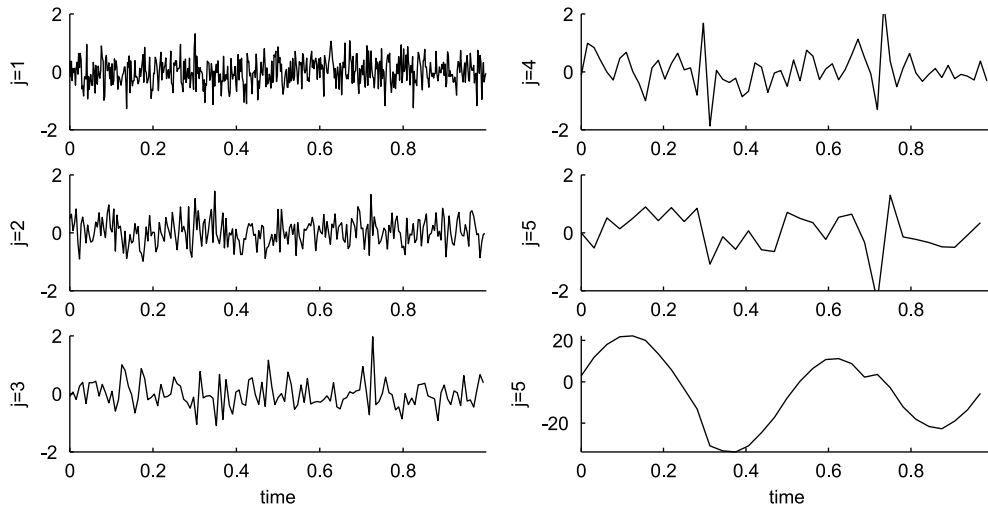
Thus if the quantity we are trying to estimate is small compared to the variance of the empirical estimator  $\hat{w}$  then by

choosing  $|c|$  smaller than one, we reduce the mean square error in estimation. Thus for  $|w|$  taking a range of values it would be better to use the shrinkage estimator. This may seem non-intuitive, but corresponds to a whole set of theory derived due to [53]. If there was an oracle that could for a given shrinkage rule inform the estimator how coefficients should be shrunk, then good estimators would be obtained. Fortunately it is not necessary to have an oracle, for large sample sizes, to expect to get an estimator with equivalent performance to using an oracle. Looking at Fig. 2 and Fig. 3 we certainly see that replacing most of the empirical wavelet coefficients by zero, should greatly improve the estimation of these coefficients.

Two special shrinkage functions are commonly used, namely the *soft* and *hard* threshold functions. The soft threshold estimator modifies the empirical estimator  $\hat{w}$  by

$$\hat{w}_\lambda^{(st)} = \begin{cases} \hat{w} - \lambda & \text{if } \hat{w} \geq \lambda \\ 0 & \text{if } |\hat{w}| < \lambda \\ \hat{w} + \lambda & \text{if } \hat{w} \leq -\lambda \end{cases} \quad (14)$$

i. e. if the observed empirical estimate is sufficiently large a set amount is removed from its magnitude, whilst if it is not sufficiently large, the estimator is set to zero. Also



Statistical Applications of Wavelets, Figure 3

The noisy wavelet coefficients of the Heavisine signal, used in the original paper by Donoho and Johnstone. The decomposition has been stopped at level five, and then the scaling coefficients have been calculated to complete the representation

popular is the hard threshold estimator, defined by:

$$\hat{w}_{\lambda}^{(ht)} = \begin{cases} \hat{w} & \text{if } |\hat{w}| \geq \lambda \\ 0 & \text{if } |\hat{w}| < \lambda \\ -\hat{w} & \text{if } \hat{w} \leq -\lambda \end{cases} \quad (15)$$

Most wavelet coefficients of sufficiently regular functions (i. e. functions in a suitable Besov space) will be very small, or even zero. One would greatly reduce the mean square error of the estimated mean function if one estimated those by zero, rather than their observed value. Thus we define the “oracle risk” in estimation as the mean square error which arises from using the best keep or kill procedure on the empirical wavelet coefficients, denoted by  $R_N$ :

$$R_N = \sum_{n=0}^{N-1} \min(w_n^2, \sigma^2) \quad (16)$$

If instead the estimator  $\hat{w}_{\sigma\lambda_N}^{(ht)}$  for  $\lambda_N = \sigma\sqrt{2\log(N)}$  (the *universal threshold*) is used, then Donoho and Johnstone showed that the mean square error of this procedure is close to the oracle risk, see [56]. Alternative methods chose a value of  $\lambda$  that is permitted to vary with the value of  $j$ .  $\sigma$  is estimated using a robust scale estimator on the first level wavelet coefficients, such as the median absolute deviation estimator.

If there is only noise present, then using the universal threshold with a hard thresholding procedure will ensure that noise only signals are estimated as zero, see [56]. [18] give a more flexible definition of a universal thresholding

procedure in terms of the probability of keeping the largest noise coefficient.

The simplest wavelet estimation method, suitable for regularly spaced, uncorrelated and Gaussian data has subsequently been extended to deal with irregularly spaced data, that is contaminated with correlated noise, and is immersed in non-Gaussian perturbations. A whole class of methods arise from transforming non-Gaussian observations, using say the Anscombe transform. Of particular note is the Haar–Fisz transform, introduced by [21]. The wavelet Haar–Fisz transform is defined by:

$$f_j^k = \begin{cases} 0 & \text{if } v_{jk} = 0 \\ \frac{w_{jk}}{\sqrt{v_{jk}}} & \text{if } v_{jk} \neq 0 \end{cases}, \quad (17)$$

and the distribution of this variable is substantially more Gaussian than the original wavelet coefficients. Fryzlewicz and Nason have established the properties of using this transform in signal estimation in a series of papers, see [21,22,23]. The denoising of correlated data was treated by [30].

Many different extensions also exist to the basic procedure for white Gaussian random variables: usually the DWT coefficients exhibit some form of structured behavior, and either clustering across time ( $l$ ) or persistence across scale ( $j$ ) is observed in the magnitude of the coefficients. Various methods have been proposed to use these observed characteristics, such as treating whole blocks of coefficients together, see [7], using Bayesian methods where the dependence of the coefficients is modeled in the

prior and other methods, such as wavelet footprints due to [19], and ‘Analytic’ denoising proposed by Olhede and Walden and later extended to two dimensions, see [43] and [42]. To observe the clustered structure of wavelet coefficients, see Fig. 2. The discontinuity in the considered signal causes wavelet coefficients to be non-zero near the discontinuities in time across all scales, and across scales at the given time location of the discontinuity. To represent the discontinuities perfectly, and avoid wavelet Gibb’s effects, all of these coefficients should be kept, thus estimating a single discontinuity without artifacts. Similar statements can be made for other features than perfect discontinuities, and the above mentioned methods all address this problem in a chosen manner.

A key characteristic of observed coefficients is the sparsity level, i. e. how many of the coefficients that are approximately zero in expectation. Choosing the threshold  $\lambda$  should be done in light of the sparsity of the data. Optimally such sparsity is learned from the data. [29] proposed the EBAYES procedure whereby at each scale the sparsity of the data is estimated and a threshold chosen using marginal likelihood.

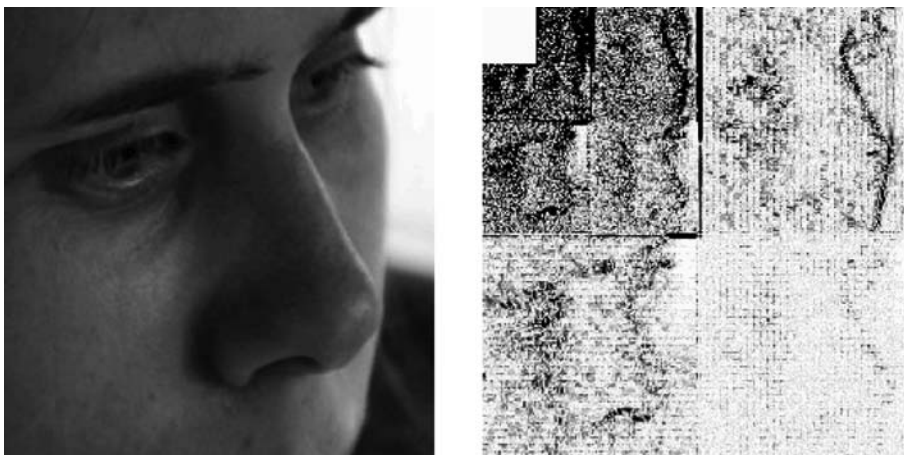
## 2-D Extensions of Wavelets

The development of statistical estimation methods for 1-D signals observed in noise is a mature field. Greater challenges lie in higher dimensional objects such as spatial, or spatiotemporal signals. In the early days when methods for extended into 2-D, the two spatial directions were treated as if the observed phenomena were two different

phenomena. This caused a number of unpalatable effects in the reconstruction of the estimated images, such as edge effects and blocking, as well as using too many coefficients to reproduce simple phenomena. To see an example of this, observe Fig. 4. The woman is mainly a smooth picture, and should only correspond to a limited set of coefficients in a compressed representation. Using separable wavelets to represent the picture, because the curve of the woman’s nose is misaligned with the decomposition, causes too many coefficients to be used. This lack of compression, leads to poor performance in the estimation.

The solution to these problems came by the development of non-separable decompositions, this leading to fewer non-zero coefficients in the decompositions, or by statistical processing of the separable decompositions, using the combined magnitude of the many coefficients to improve estimation. The main methods of note are ridgelets and curvelets proposed by [9,52], the dual-tree complex wavelet transform, proposed by Kingsbury (see [48]), and bandelets (see [33]). The implementation of several of these transforms is slightly more complicated than the 1-D wavelet transform, but all correspond to linear operations on the observed signal. This means that the statistical properties of the transform coefficients can be determined by direct (if sometimes numerical) calculations. Again, thresholding and inverse transformation forms the basis for most estimation procedures, much in the same manner as the 1-D estimation procedures.

Open areas of research in this general setting, is the development of higher dimensional processing, and suitable transformations for discrete analysis on the sphere.



Statistical Applications of Wavelets, Figure 4

A picture of a woman (*left*), and the absolute value of the separable 2-D wavelet coefficients of the picture (*right*). Observe that the nose of the woman takes too many coefficients to record, despite the curve of the nose corresponding to a smooth curve. This illustrates the lack of compression of the 2-D separable wavelet transform



### Covariance Estimation Using Wavelets

Wavelets have been used to estimate properties of zero-mean stochastic processes, characterized by their dependence structure alone. One of the fundamental aspects is noting classes of stochastic processes for which the transform provides a simplification of the description of the signal. One set of developments has defined a set of locally stationary processes using wavelet filters as the building block, so called Locally Stationary Wavelet Processes, see [41]. The estimation of the fundamental representation of this process, which is very much like a time varying spectrum, was addressed by Nason and his coauthors, and has subsequently been improved upon by Fryzlewicz and Nason, see [23]. The model has been used to classify sleep states of babies, and has been used in modeling financial times series, see [5,20]. This corresponds to the most natural modification of existing theory for locally stationary time series from local Fourier to wavelet bases.

The usage of wavelets for estimating time series that when differenced are stationary has also seen substantial investigation. By using wavelets the number of differencing steps needs often not be determined, which is a clear advantage. Likelihood theory has been extended from the Fourier domain to the wavelet domain for such time series, see [39], and various procedures for estimating the decay of long memory processes have been adapted for wavelet coefficients, see [38]. Characterization of sets of time series that are jointly stationary when differenced, has also been developed, see [49,50,51,57]. A fundamental method of analysis is to characterize the covariances of the process associated with different time-scales, this yielding a multiresolution analysis of the stochastic process, both considering the auto-covariance of a single process, and cross-covariances between several processes.

### Future Directions

The usage of the wavelet transform in 1-D is a mature field. Current developments are mainly in the application and adaption of existing methods to novel applications. The development of higher dimensional transforms has also slowed down considerably. The modern thrust and drive is instead found in using several concepts which the development of analysis methods from the wavelet transform were founded upon. The key two notions for these developments are sparsity and incoherence. For many signals of interest, their wavelet coefficient representation was compressed. This means the sparsity could be used to substantially improve the estimation procedure. The area of 'compressed sensing' is based on using the compression of a signal in a known decomposition, to improve the estima-

tion of the signal, see [14]. Surprisingly the limitations of Nyquist sampling has been circumvented, and near miraculous reconstruction performance obtained. As the data deluge that faces modern signal processing will increase, this stands as a very interesting area of future development.

### Bibliography

1. Antoniadis A, Paparoditis E, Sapatinas T (2006) A functional wavelet-kernel approach for time series prediction. *J Roy Stat Soc B* 68:837–857
2. Antoniadis A, Sapatinas T (2003) Wavelet methods for continuous-time prediction using Hilbert-valued autoregressive processes. *J Multivar Anal* 87:133–158
3. Averkamp R, Houdre C (1998) Some distributional properties of the continuous wavelet transform of random processes. *IEEE Trans Info Theory* 44:1111–1124
4. Averkamp R, Houdre C (2000) A note on the discrete wavelet transform of second-order processes. *IEEE Trans Info Theory* 46:1673–1676
5. Bellegem SV, Fryzlewicz P, von Sachs R (2003) A wavelet-based model for forecasting non-stationary processes. *Inst Phys Conf Ser* 173:955–958
6. Buresti G, Lombardi G, Bellazzini J (2004) On the analysis of fluctuating velocity signals through methods based on the wavelet and hilbert transforms. *Chaos Solitons Fractals* 20:149–158
7. Cai T, Silverman BW (2001) Incorporating information on neighbouring coefficients into wavelet estimation. *Sankhyā Ser B* 63:127–148
8. Cambanis S, Houdre C (1995) On the continuous wavelet transform of second-order random processes. *IEEE Trans Info Theory* 41:628–642
9. Candès E, Donoho DL (1999) Ridgelets: a key to higher-dimensional intermittency? *Philos Trans: Math Phys Eng Sci* 357:2495–2509
10. Daubechies I (1988) Orthonormal bases of compactly supported wavelets. *Comm Pure Appl Math* 41:909–996
11. Daubechies I (1992) Ten lectures on wavelets. SIAM, Philadelphia
12. Daubechies I, Sweldens W (1998) Factoring wavelet transforms into lifting steps. *J Fourier Anal Appl* 4:247–269
13. Delprat N, Escudié B, Guillemain P, Kronland-Martinet R, Tchamitchian P, Torresani B (1992) Asymptotic wavelet and gabor analysis: extraction of instantaneous frequencies. *IEEE Trans Inform Theory* 38:644–64
14. Donoho DL (2006) Compressed sensing. *IEEE Trans Inf Theory* 52:1289–1306
15. Donoho DL, Johnstone IM (1994) Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81:425–455
16. Donoho DL, Johnstone IM (1995) Adapting to unknown smoothness via wavelet shrinkage. *J Am Stat Assoc* 90:1200–1224
17. Donoho DL, Johnstone IM, Kerkyacharian G, Picard D (1995) Wavelet shrinkage – asymptotia. *J Roy Stat Soc B* 57:301–337
18. Downie TR, Silverman BW (1998) The discrete multiple wavelet transform and threshold methods. *IEEE Trans Signal Process* 46:2558–2561

19. Dragotti PL, Vetterli M (2003) Wavelet footprints: theory, algorithms, and applications. *IEEE Trans Signal Process* 51:1306–1323
20. Fryzlewicz P, Bellegem SV, von Sachs R (2003) Forecasting non-stationary time series by wavelet process modelling. *Ann Inst Stat Math* 55:737–764
21. Fryzlewicz P, Nason GP (2004a) A Haar–Fisz algorithm for poisson intensity estimation. *J Comp Graph Stat* 13:621–638
22. Fryzlewicz P, Nason GP (2004b) A Haar–Fisz algorithm for poisson intensity estimation. *J Comp Graph Stat* 13:621–638
23. Fryzlewicz P, Nason GP (2006) Haar–Fisz estimation of evolutionary wavelet spectra. *J Roy Stats Soc B* 68:611–634
24. Grossman A, Morlet J (1984) Decomposition of Hardy functions in square integrable wavelets of constant shape. *Siam J Math Anal* 15:723–736
25. Guérin C-A (2000) Wavelet analysis and covariance structure of some classes of non-stationary processes. *J Fourier Anal Appl* 6:403–425
26. Haar A (1910) On the theory of orthogonal function systems. *Math Ann* 69:331–371
27. Holschneider M (1995) Wavelets: an analysis tool. Oxford University Press, Oxford
28. Jaffard S, Melot S (2005) Wavelet analysis of fractal boundaries. *Commun Math Phys* 258:153–159
29. Johnstone I, Silverman BW (2005) Empirical Bayes selection of wavelet thresholds. *Ann Stat* 33:1700–1752
30. Johnstone IM, Silverman BW (1997) Wavelet threshold estimators for data with correlated noise. *J Roy Stat Soc Ser B* 59:319–351
31. Kim IK, Kim YY (2005) Damage size estimation by the continuous wavelet ridge analysis of dispersive bending waves in a beam. *J Sound Vib* 287:707–722
32. Krim H, Pesquet J-C (1995) Multiresolution analysis of a class of nonstationary processes. *IEEE Trans Inf Theory* 41:1010–1020
33. le Pennec E, Mallat S (2005) Sparse geometric image representations. *IEEE Trans Image Process* 14:423–438
34. Li T-H, Oh H-S (2002) Wavelet spectrum and its characterization property for random processes. *IEEE Trans Inf Theory* 48:2922–2937
35. Lilly JM, Gascard J-C (2006) Wavelet ridge diagnosis of time-varying elliptical signals with application to an oceanic eddy. *Nonlinear Process Geophys* 13:467–483
36. Loève M (1945) Sur le fonctions aléatoire de second ordre. *Rev Sci* 83:297–303
37. Mallat S (1999) A Wavelet Tour of Signal Processing, 2nd edn. Academic Press, New York
38. McCoy EJ, Walden AT (1996) Wavelet analysis and synthesis of stationary long-memory processes. *J Comp Graph Stat* 5:26–56
39. Moulines E, Roueff F, Taqqu MS (2007) On the spectral density of the wavelet coefficients of long-memory time series with application to the log-regression estimation of the memory parameter. *J Time Ser Anal* 28:155–187
40. Nason GP, von Sachs R (1999) Wavelets in time-series analysis. *Philos Trans Math Phys Eng Sci* 357:2511–2526
41. Nason GP, von Sachs R, Kroisandt G (2000) Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *J Roy Stat Soc B* 62:271–292
42. Olhede S (2007) Hyperanalytic denoising. *IEEE Trans Image Process* 16:1522–1537
43. Olhede S, Walden AT (2004a) ‘Analytic’ denoising. *Biometrika* 91:955–973
44. Olhede SC, Walden AT (2004b) The Hilbert spectrum via wavelet projections. *Proc Roy Soc Lond A* 460:955–975
45. Olhede SC, Walden AT (2005) Wavelet denoising for signals in quadrature. *Integr Computer-Aided Eng* 12:109–117
46. Percival DB, Walden AT (2000) Wavelet methods for time series analysis. Cambridge University Press, Cambridge
47. Schuster A (1898) On the investigation of hidden periodicities with application to a supposed 26-day period of meteorological phenomena. *Terr Mag Atmos Elect* 3:13–41
48. Selesnick IW, Baraniuk RG, Kingsbury NG (2005) The dual-tree complex wavelet transform. *IEEE Signal Process Mag* 22(6):123–151
49. Serroukh A, Walden AT (2000a) Wavelet scale analysis of bivariate time series i: motivation and estimation. *J Nonparametr Stat* 13:1–36
50. Serroukh A, Walden AT (2000b) Wavelet scale analysis of bivariate time series ii: statistical properties for linear processes. *J Nonparametr Stat* 13:37–56
51. Serroukh A, Walden AT, Percival DB (2000) Statistical properties and uses of the wavelet variance estimator for the scale analysis of time series. *J Am Stat Assoc* 95:184–196
52. Starck JL, Candès EJ, Donoho DL (2002) The curvelet transform for image denoising. *IEEE Trans Image Process* 11:670–84
53. Stein C (1981) Estimation of the mean of a multivariate normal distribution. *Ann Stat* 9:1135–1151
54. Torrence C, Compo GP (1998) A practical guide to wavelet analysis. *Bull Am Meteorol Soc* 79:61–78
55. Tu C-L, Hwang W-L, Ho J (2005) Analysis of singularities from modulus maxima of complex wavelets. *IEEE Trans Inf Theory* 51:1049–1062
56. Wasserman L (2006) All of nonparametric statistics. Springer, New York
57. Whitcher B, Guttorp P, Percival DB (2000) Wavelet analysis of covariance with application to atmospheric time series. *J Geophys Res Atmos* 105(14):14941–14962
58. Wolfe PJ, Godsill SJ, Ng WJ (2004) Bayesian variable selection and regularisation for time-frequency surface estimation. *J Roy Stat Soc Ser B* 66:575–589
59. Zhang Z, Ren Z, Huang W (2003) A novel detection method of motor broken rotor bars based on wavelet ridge. *IEEE Trans Energy Convers* 18:417–423

---

## Statistical and Non-linear Physics, Introduction to

M. CRISTINA MARCHETTI

Physics Department, Syracuse University, Syracuse, USA

The field of Statistical and Nonlinear Physics combines the venerable subject of statistical mechanics with the newer area of nonlinear science to create a highly interdisciplinary and exciting area of research. It has its roots in equilibrium statistical physics, but it has evolved to encompass and emphasize nonequilibrium and dynamical phenomena. It is a field in rapid evolution, with a constantly changing focus. It overlaps naturally with many

other disciplines, including fluid dynamics, computational physics, biological physics, condensed matter physics and polymer physics.

Statistical physics aims to describe the large-scale collective behavior of systems composed of a large number of interacting units or degrees of freedom. Starting with a microscopic model, one performs various types of “coarse-graining” to describe phenomena that occur on length and time scales large compared to microscopic ones, such as the size of the particles or the characteristic interaction times. At this large scale the system exhibits collective or emergent behavior that is far richer than that of the individual units. One of the hallmarks of collective behavior is phase transitions. An everyday example is the change from liquid to solid that occurs upon tuning a parameter such as temperature or pressure. Another familiar example arises in the study of the properties of magnetic materials and is highlighted in the article [► Complex Systems and Emergent Phenomena](#), which is a good starting point for the reader of this section.

When going from the deterministic Hamiltonian description of a system of many interacting units to the large scale coarse-grained description of the same system in terms of a continuum or hydrodynamic theory, an element of randomness is naturally introduced into the problem as one loses the invariance under time reversal that was present in the Hamiltonian description. At the same time randomness on the other hand is an intrinsic property of the dynamics of individual nonlinear systems that are known to often exhibit sensitive dependence to initial conditions and chaotic behavior. There is in fact a deep connection between the randomness of chaotic systems and the irreversible transport properties of extended systems (see [► Chaotic Dynamics in Nonequilibrium Statistical Mechanics](#)). This connection highlights the unity of the two areas of statistical and nonlinear science.

Rather than presenting an exhaustive review of the field of statistical and nonlinear physics, the articles in this section focus on nonequilibrium phenomena that are the subject of current research. Even then, only a fraction of the problems and physical systems that are studied using the tools and ideas of statistical and nonlinear physics are described here. Many of the others can be found in articles throughout the rest of the Encyclopedia.

The topics highlighted in this section of the Encyclopedia may at a first glance seem disparate, but are unified by the ideas that are used to study them. Principal among those are the notions of scaling and universality. The concept of universality, which has been around for some time, has its roots in the study of phase transitions and critical phenomena. Near a phase transition, the system is univer-

sal in that its behavior at large scales does not depend on the microscopic physics. The systems and physical problems described in this section of the Encyclopedia may be different in the details, but generally involve the onset of collective or emergent behavior with “universal” properties.

The ideas of scaling universality have had a crucial impact in the study of long-standing problems in nonlinear physics, such as pattern formation and turbulence (see [► Non-linear Fluid Flow, Pattern Formation, Mixing and Turbulence](#)). The same tools and ideas have more recently been applied to the study of the complex dynamics of neuronal systems (see [► Neuronal Dynamics](#)), lasers (see [► Noise and Stability in Modelocked Soliton Lasers](#)), and even the physics of economics (see [► Economics, Statistical Mechanics Approach to](#)). Rich chaotic behavior has also been discovered in quantum systems (see [► Quantum Chaos](#)).

Recently, a new field has emerged that uses many of the ideas of statistical mechanics to explore problems in “soft” condensed matter physics. Soft materials are literally materials that yield easily to mechanical deformations. They encompass an extremely wide range of systems, from complex fluids such as colloidal suspensions, liquid crystals and polymers (see [► Polymer Physics](#)) to soft elastic gels (see [► Anisotropic Networks, Elastomers and Gels](#)), and granular matter in both its fluidized (see [► Granular Flows](#)) and its jammed (see [► Jamming of Granular Matter](#)) states. The jamming or structural arrest of granular matter when the density reaches a critical value bears striking similarities with the glass transition. Jammed granular matter and glassy systems share many of the intriguing dynamical properties discussed in [► Glasses and Aging, A Statistical Mechanics Perspective on](#). Jamming also provides an example of the nonequilibrium phase transitions that are ubiquitous in the systems studied by modern statistical physics. Nonequilibrium phase transitions do occur in “pure” systems, i. e., ones without any extrinsic disorder, as a result of the dynamical constraints arising from interactions, but are even more common in systems with quenched disorder, induced for instance by impurities and materials defects (see [► Disordered Elastic Media](#) for a general perspective on such systems). Notable examples include the depinning transition that occurs in driven extended systems such as vortex lattices in type-II superconductors under the action of a uniform external driving force (see [► Collective Transport and Depinning](#)) and the slip-stick motion that characterizes the physics of friction and tectonic motion (see [► Jerky Motion in Slowly Driven Magnetic and Earthquake Fault Systems, Physics of](#)). Living matter, such as the cytoskeleton of eukaryotic cell or

even the whole cell, has also been described as a novel form of nonequilibrium soft matter (see ► [Cytoskeleton and Cell Motility](#)).

The complexity of the systems that are studied by statistical physicists has naturally led to the development of sophisticated numerical methods. Principal among those is the technique of Monte Carlo simulations (see ► [Monte Carlo Simulations in Statistical Physics](#)). More recently physicists have also begun to adapt ideas and algorithms from computer science (see ► [Optimization Problems and Algorithms from Computer Science](#)).

Novel experimental techniques have recently allowed us to begin to study the properties of soft and living matter on length and time scales that require a qualitative new way of thinking and a modification of the familiar tools from statistical physics (see ► [Fluctuation Theorems, Brownian Motors and Thermodynamics of Small Systems](#)). These range from optical techniques for study of the properties of individual biomolecules (see ► [Protein Mechanics at the Single-Molecule Level](#)) to microfluidics devices for the study of the rheological and mechanical properties of complex materials and cells on novel time and length scales (see ► [Microfluidics](#)).

Yet another class of problems that has caught the attention of statistical physicists lately is the study of networks (see ► [Networks: Structure and Dynamics](#)). Examples of networks are as disparate as the world-wide web, human societies, languages and their structure, and power grids. An especially important class of networks occurs in biology. It consists of the cell signaling networks that control many of the functions of living cells (see ► [Cell Biology: Networks, Regulation and Pathways](#)).

Finally, complex emergent behavior is also ubiquitous in quantum matter, where the interactions and fluctuations are controlled by quantum mechanics, as described in the article ► [Ultracold Atomic Gases: Novel States of Matter](#).

## Statistics with Imprecise Data

MARÍA ÁNGELES GIL<sup>1</sup>, OLGIERD HRYNIEWICZ<sup>2</sup>

<sup>1</sup> Faculty of Sciences, University of Oviedo, Oviedo, Spain

<sup>2</sup> Systems Research Institute, Warsaw, Poland

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Mathematical Modeling of Imprecise Data](#)

[Fuzzy Random Variables](#)

[Statistical Analysis of Fuzzy Data Corresponding to Fuzzy Perceptions of Existing Real-Valued Data](#)

[Statistical Analysis of Existing Fuzzy-Valued Data](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Fuzzy estimators** Estimators of “parameters” of probability distributions, or other characteristics, of random variables/fuzzy random variables (such as e. g. the expected value/the fuzzy expected value) when statistical data are imprecise and are described by means of fuzzy sets.

**Fuzzy random variable** Random element whose observed values are described by fuzzy sets.

**Fuzzy set** Generalization of a classical notion of a set. In contrast to the case of a classical set, each element  $x$  of a fuzzy set may belong to it to a degree described by the so-called membership function  $\mu(x)$ . Thus, the fuzzy set may be defined as a set of ordered pairs  $(x, \mu(x))$ , where  $x$  belongs to a set  $\mathbb{X}$  called the universe of discourse or referential. Alternatively, the fuzzy set can be identified with its membership function (in the same way that a classical set can be identified with its indicator function).

**Fuzzy statistical tests** Statistical tests used for the verification of hypotheses about the values of “parameters” of probability distributions, or other characteristics, of random variables/fuzzy random variables when statistical data are imprecise and are described by fuzzy sets.

**Fuzzy statistics** Generalization of traditional statistics that allows to handle imprecise data described in terms of fuzzy sets.

**Imprecise data** Data that cannot be described by either real numbers or vectors with real-valued components.

**Random sets** Random elements whose observed values are sets (e. g. intervals, subsets of a plane).

### Definition of the Subject

Traditional statistics deals usually with precisely defined data. The source of these data may be of different nature. Usually the data are collected from observations of random experiments, i. e. experiments whose performance leads to an outcome which cannot be predicted in advance with one hundred percent sureness. When the knowledge about the nature of these random events is available we can use the methods of mathematical statistics and draw conclusions about the source of available data. Uncertainty being intrinsic to random outcomes/events is properly de-



scribed by using the formalism of the mathematical theory of probability, and generally it is attributed to *future* observations of these outcomes/events.

Traditionally, statistical experimental data are described by real numbers or by vectors whose components are real numbers. These numbers are either observed directly as results of measurements (e.g. height and weight of a person) or correspond to observed counts of certain categories representing labeled events (e.g. a gender of that person). However, in real life applications the results of measurements are never precise. The precision of every measurement is limited by the precision of a measuring device. In the majority of practical applications this lack of precision may be neglected, and for statistical analysis of available data we can use traditional methods of statistics. If the measurement error cannot be neglected statistical analysis of interval data is recommended by specialists in metrology. However, when statistical data are presented by human beings we need more sophisticated methods for the description of their lack of precision. Therefore, there is a practical need to generalize traditional statistical methods in order to make them applicable for handling imprecise data, e.g. the data described by statements expressed by using a plain language (the so-called “linguistic data”).

## Introduction

There is a common agreement that uncertainty characterized by randomness shall be described by considering the theory of probability. Thus, mathematical statistics is a proper tool for dealing with data generated by random experiments and described by precisely defined numbers. However, there also exist other types of uncertainty which are related to vagueness, imprecision, existence of only partial information about experimental outcomes/events of interest, etc. In contrast to randomness, uncertainties of such types are attributed rather to *current* perceptions/observations. It has to be noted that the mathematical modeling of all these types of uncertainty which are different from simple randomness is still the subject of controversies. A good overview of the related problems can be found in the paper by P. Walley [36]. In this article we confine ourselves to the case when randomness is observed together with vagueness understood as the lack of precision.

Specialists in measurement theory recognize different types of uncertainty. For instance, in the ISO/IEC Guide [18] it is recommended to distinguish between uncertainty related to pure randomness and uncertainty of other nature, such as a lack of precision. However, the lack of precision of statistical data is usually omitted in the sta-

tistical analysis of measurements. Consider, for example, the analysis of results of measurements coming from a digital meter. The results of measurements coming from such a meter are always rounded in order to have their representation by a certain number of digits. When we observe a displayed result of a measurement we never know what is the actual value of the measured quantity. What is more important, and often overlooked by statisticians, we may not increase our knowledge about that value by averaging the results of repeated measurements, as it is frequently recommended by statisticians. Consider, for example, a series of measurements showing exactly the same result. Having such results we are in principle not able to distinguish between two fundamentally different cases when the measured quantity is constant and its value belongs to the range of rounding or when it is varying from measurement to measurement with values belonging to that range.

The situation described in the previous paragraph is relatively simple as the range of possible actual values corresponding to the observed result of a measurement is usually precisely defined. There exist, however, situations when either this range is not precisely known or values in the range are not equally compatible with the available data/information/events. We face such cases when results of measurements or classifications are evaluated by humans (e.g., by evaluating indications of an analog meter or classifying individuals in accordance with their height, and so on). The result of a measurement is even more imprecise when it has been obtained without the usage of any meter (e.g., when we visually evaluate the distance between two points); in such situations statistical data may consist of imprecise statements like “around 5 meters”, “more or less between 5 and 10 seconds”, etc. The classification of a person in accordance with his/her height leads often to imprecise assessments, like “very short”, “rather tall”, etc. A similar situation occurs when we deal with retrospective data recalled by human beings; for example, in reliability analysis of field lifetime data we may face situations when failure times are reported imprecisely using statements like ‘the failure occurred about one month ago’, etc. In all these cases statistical data consist of *imprecise perceptions* of actual real values.

In the previous paragraphs we considered situations when actual values of measured quantities exist, but they are imprecisely perceived. There exist, however, situations when we have to analyze statistical data that represent imprecisely defined concepts. Take for example the color of human hair described using categories such as “blond” or “dark blond”; it is obvious that the border between these two categories is vague; one can try to establish a precise



border between these two categories in terms of the results of precise measurements of the spectrum of reflected light, but such attempts seem to be practically senseless. We face a similar situation in classifying a client of a bank office in accordance with his/her degree of aversion to investment which leads to imprecise assessments, like “very low”, “moderate”, etc. In both these cases we could try to collect precise statistical data, e. g., by either asking a respondent to a questionnaire to indicate exactly one choice or coding it numerically. However, it seems to be more prudent, natural and informative to expect imprecise answers to questions pertained to vague notions. If we do so, we may face imprecise statistical data for further analysis. In all these cases statistical data consist of *imprecise actual values* themselves.

There also exists another source of imprecision while dealing with statistical data. For example, in reliability lifetime tests we perform tests in more severe “over-stress” conditions, and then we try to recalculate test results to conditions which are considered “normal”. For this recalculation we can utilize some partial knowledge about possible values of stress-dependent recalculation coefficients. In such a case we have originally precise lifetime data, but after recalculation these data become imprecise.

In all considered cases the lack of precision can be appropriately described by using fuzzy sets introduced by Lotfi A Zadeh. In the second section of the article we recall some basic definitions related to fuzzy sets. We will present the fuzzy sets methodology as a useful tool for the description of imprecise data. When fuzzy lack of precision is mixed with randomness, either in the sense that available fuzzy data are supposed to come from the perception of real- or vectorial-valued data generated by a random mechanism, or in the sense of they being directly generated by a random mechanism, a convenient tool to use is that of the notion of a fuzzy random variable. This notion is introduced in the third section of this article. The interpretation of the fuzzy random variable depends upon the type of observed fuzzy data and events. In the paper we distinguish between the two types of data which have been described in this section. We start with the description of statistical methods which are useful for the analysis of fuzzy perceptions of existing precise values. Then, we present statistical methods which are useful in the analysis of intrinsically fuzzy-valued data.

## Mathematical Modeling of Imprecise Data

There exist competitive methods for the description of vagueness. For example, some statisticians claim that the theory of subjective probability is sufficient for the descrip-

tion of all types of uncertainty. However, many other researchers have shown examples of situations when the application of the classical theory of probability is not sufficient for modeling these situations. Therefore, other formalisms have been proposed for the description of vagueness/imprecision. One of those formalisms, namely the theory of *fuzzy sets* proposed by Lotfi A. Zadeh [37], has been slowly but widely accepted as a good methodology for the description of imprecise data, both from practical and theoretical (see, e. g., the paper by Terán [33]) points of view.

The basic concept of the theory of fuzzy sets is the universe of discourse or referential  $\mathbb{X}$  which may be understood as the set of all possible (feasible) elements that are relevant for the description of a certain concept (quantity). Mathematically speaking a *fuzzy subset*  $A$  of a set  $\mathbb{X} \neq \emptyset$  (or a fuzzy set, for short) is a map  $A : \mathbb{X} \rightarrow [0, 1]$ , where  $A(x)$  can be interpreted as the degree of compatibility of  $x$  with the ill-defined property characterizing  $A$ , or degree of truth of the assertion “ $x$  is  $A$ ”, or degree of membership of  $x$  to  $A$ . Equivalent, but more intuitive for some purposes, a fuzzy set  $A$  of  $\mathbb{X}$  can be defined as a set of ordered pairs  $\{(x, \mu_A)\}$ , where  $x \in \mathbb{X}$  and  $\mu_A : \mathbb{X} \rightarrow [0, 1]$  is the so-called *membership function* of  $A$ . In other words, a fuzzy set can be identified with its membership function, in the same way that a classical set can be identified with its indicator function. In what follows, we will consider indistinctly  $A$  or  $\mu_A$  to denote and refer to a fuzzy subset.

Unfortunately, there is no one generally accepted methodology for the construction of membership functions. The majority of researchers assume that the membership function  $\mu_A(x)$  is a purely subjective function provided by a person who describes his/her perception of a certain phenomenon or quantity. Some authors provide practical methods for the construction of the membership function when it is interpreted in terms of the theory of possibility as the possibility distribution (see [8] for more information). Some other authors, e.g. Bandemer and Näther [1] or Viertl [35], present methods which may be used for the construction of membership functions in a more objective way from measurements of physical quantities. Anyway, our purpose is not entering a discussion here about this point.

In the analysis of imprecise data we are usually interested in the description of interesting phenomena by numbers. For this purpose we can use the concept of a fuzzy number defined as follows (see [5]):

**Definition 1** The fuzzy subset  $A$  of the space of real numbers  $\mathbb{R}$ , with the membership function  $\mu_A : \mathbb{R} \rightarrow [0, 1]$ , is a **fuzzy number** if and only if

- (a)  $A$  is normal, i. e. there exists at least an element  $x_0 \in \mathbb{R}$  such that  $\mu_A(x_0) = 1$ ;
- (b)  $A$  is fuzzy convex, i. e.,  $\mu_A(\lambda x_1 + (1-\lambda)x_2) \geq \mu_A(x_1) \wedge \mu_A(x_2)$ , for all  $x_1, x_2 \in \mathbb{R}$ , and  $\lambda \in [0, 1]$ ;
- (c)  $\mu_A$  is upper semicontinuous;
- (d) the support set,  $\text{supp } A = \{x \in \mathbb{R} : \mu_A(x) > 0\}$ , is bounded (that is, its closure  $\text{cl}(\text{supp } A)$  is compact).

It is easily seen that if  $A$  is a fuzzy number then its membership function can be expressed as follows:

$$\mu_A(x) = \begin{cases} 0 & \text{for } x < a_1 \\ r_{l_A}(x) & \text{for } a_1 \leq x < a_2 \\ 1 & \text{for } a_2 \leq x \leq a_3 \\ r_{u_A}(x) & \text{for } a_3 < x \leq a_4 \\ 0 & \text{for } x > a_4, \end{cases} \quad (1)$$

where  $a_1, a_2, a_3, a_4 \in \mathbb{R}$ ,  $a_1 \leq a_2 \leq a_3 \leq a_4$ ,  $r_{l_A} : [a_1, a_2] \rightarrow [0, 1]$  is a nondecreasing upper semicontinuous function, and  $r_{u_A} : [a_3, a_4] \rightarrow [0, 1]$  is a nonincreasing upper semicontinuous function. Functions  $r_{l_A}$  and  $r_{u_A}$  are called sometimes the left and the right “arms” (or “sides”) of the fuzzy number, respectively.

A useful notion for dealing with a fuzzy number is the so-called  $\alpha$ -level set, also known as the  $\alpha$ -cut. For  $\alpha \in (0, 1]$  the  $\alpha$ -level set of the fuzzy number  $A$  is the ordinary (non-fuzzy) set defined as

$$A_\alpha = \{x \in \mathbb{R} : \mu_A(x) \geq \alpha\} \quad (2)$$

and the 0-level set is usually intended to be given by  $A_0 = \text{cl}(\text{supp } A)$ . The family  $\{A_\alpha : \alpha \in [0, 1]\}$  is a set representation of the fuzzy number  $A$ . According to the resolution identity proposed by Zadeh, we can represent the membership function as:

$$\mu_A(x) = \sup\{\alpha \cdot \mathbf{1}_{A_\alpha}(x) : \alpha \in (0, 1]\}, \quad (3)$$

where  $\mathbf{1}_{A_\alpha}(x)$  denotes the indicator function of  $A_\alpha$ .

On the basis of the notion of  $\alpha$ -level, a fuzzy number  $A$  can be viewed as a fuzzy subset of  $\mathbb{R}$  such that its  $\alpha$ -level sets are nonempty compact and convex sets of  $\mathbb{R}$ , that is, nonempty compact intervals. Hence, for each  $\alpha \in [0, 1]$  we have that  $A_\alpha = [A_L(\alpha), A_U(\alpha)]$ , where

$$\begin{aligned} A_L(\alpha) &= \inf\{x \in \mathbb{R} : \mu_A(x) \geq \alpha\}, \\ A_U(\alpha) &= \sup\{x \in \mathbb{R} : \mu_A(x) \geq \alpha\}. \end{aligned} \quad (4)$$

If the sides of the fuzzy number  $A$  are strictly monotonic functions, then from Eq. (1) one can see easily that  $A_L(\alpha)$  and  $A_U(\alpha)$  are inverse functions of  $r_{l_A}$  and  $r_{u_A}$ , respectively.

In statistical analysis of random data we use functions (and, in particular, operations) of the observed random samples. These functions, called statistics, can be also defined for fuzzy random data. Their membership functions can be derived by using Zadeh's *extension principle*. This principle has the following formulation [38]:

Let  $\mathbb{X}$  be a Cartesian product of universes  $\mathbb{X} = \mathbb{X}_1 \times \dots \times \mathbb{X}_r$ , and  $A_1, \dots, A_r$  be  $r$  fuzzy subsets of  $\mathbb{X}_1, \dots, \mathbb{X}_r$ , respectively. Let  $f$  be a mapping from  $\mathbb{X} = \mathbb{X}_1 \times \dots \times \mathbb{X}_r$  to a universe  $\mathbb{Y}$  such that  $y = f(x_1, \dots, x_r)$ . The extension principle allows us to induce from the  $r$  fuzzy sets  $A_i$  a fuzzy set  $B$  on  $\mathbb{Y}$  through  $f$ ,  $B = f(A_1, \dots, A_r)$  such that

$$\mu_B(y) = \begin{cases} \sup_{x_1, \dots, x_r | y = f(x_1, \dots, x_r)} \min\{\mu_{A_1}(x_1), \dots, \mu_{A_r}(x_r)\} & \text{if } f^{-1}(y) \neq \emptyset \\ 0 & \text{if } f^{-1}(y) = \emptyset \end{cases} \quad (5)$$

In case of  $A_i$ ,  $i = 1, \dots, n$  being fuzzy numbers, and  $f$  being either an injective or a continuous function, then for each  $\alpha \in [0, 1]$  the  $\alpha$ -level of  $B = f(A_1, \dots, A_r)$  can be shown to be equal to  $B_\alpha = (f(A_1, \dots, A_r))_\alpha$  with

$$\begin{aligned} (f(A_1, \dots, A_r))_L(\alpha) &= \min_{(x_1, \dots, x_r) \in (A_1)_\alpha \times \dots \times (A_r)_\alpha} f(x_1, \dots, x_r), \\ (f(A_1, \dots, A_r))_U(\alpha) &= \max_{(x_1, \dots, x_r) \in (A_1)_\alpha \times \dots \times (A_r)_\alpha} f(x_1, \dots, x_r). \end{aligned} \quad (6)$$

Thus, the application of the extension principle for the calculation of the membership function of  $y = f(x_1, \dots, x_r)$  is equivalent to the application of the interval arithmetics on  $\alpha$ -level sets of the arguments of this function (see [29] as a basis for the proof). For instance, if  $A$  and  $B$  are fuzzy numbers, then,

$$\begin{aligned} (A + B)_L(\alpha) &= A_L(\alpha) + B_L(\alpha), \\ (A + B)_U(\alpha) &= A_U(\alpha) + B_U(\alpha), \\ (\lambda \cdot A)_L(\alpha) &= \begin{cases} \lambda \cdot A_L(\alpha) & \text{if } \lambda \geq 0 \\ \lambda \cdot A_U(\alpha) & \text{if } \lambda < 0 \end{cases} \\ (\lambda \cdot A)_U(\alpha) &= \begin{cases} \lambda \cdot A_U(\alpha) & \text{if } \lambda \geq 0 \\ \lambda \cdot A_L(\alpha) & \text{if } \lambda < 0 \end{cases} \end{aligned}$$

for any  $\lambda \in \mathbb{R}$ , and  $\alpha \in [0, 1]$ .

### Fuzzy Random Variables

Uncertainty, understood as randomness, is well described in Probability Theory. The concept of a random variable,

which is basic in this theory, is well-known and its definition does not need to be recalled in this article.

However, when we observe random experimental data which are imprecise, a useful tool to model either the imprecise perception of values coming from real-valued random variables or the random mechanisms generating directly these imprecise data is the one associated with the so-called concept of fuzzy random variables. Actually, we can consider two different approaches to the concept of fuzzy random variable; the motivation for these approaches and the situations they apply to are different, but the formalization of the second notion and the associated statistical methodology can be applied to the first one.

Historically, the first widely accepted definition of the fuzzy random variable was proposed by Kwakernaak [21,22]. Kruse [19] proposed an interpretation of this notion, and according to this interpretation a fuzzy random variable  $Z$  may be considered as a fuzzy perception of an unknown true real-valued random variable  $Z_0$  associated with a random experiment, and referred to as ‘the original’ of  $Z$ . Below, we recall the version of this definition elaborated by Kruse and Meyer [20].

**Definition 2 (Kruse and Meyer [20])** Let  $(\Omega, \mathcal{A}, P)$  be a probability space, where  $\Omega$  is the set of all possible outcomes of a random experiment,  $\mathcal{A}$  is a  $\sigma$ -field of subsets of  $\Omega$  (the set of all possible events of interest), and  $P$  is a probability measure associated with  $(\Omega, \mathcal{A})$ .

A mapping  $X : \Omega \rightarrow \mathcal{F}_c(\mathbb{R})$ , where  $\mathcal{F}_c(\mathbb{R})$  is the space of all fuzzy numbers, is called a **fuzzy random variable** if it satisfies the following properties:

- i)  $\{X_\alpha(\omega) : \alpha \in [0, 1]\}$ , where  $X_\alpha(\omega) = (X(\omega))_\alpha$  is a set representation of  $X(\omega)$  for all  $\omega \in \Omega$ ;
- ii) for each  $\alpha \in [0, 1]$  both  $X_\alpha^L : \Omega \rightarrow \mathbb{R}$  and  $X_\alpha^U : \Omega \rightarrow \mathbb{R}$ , with  $X_\alpha^L(\omega) = \inf X_\alpha(\omega)$  and  $X_\alpha^U(\omega) = \sup X_\alpha(\omega)$ , are usual real-valued random variables associated with  $(\Omega, \mathcal{A}, P)$ .

Values of a fuzzy random variable in Definition 2 have been conceived to model fuzzy perceptions of existing real-valued values (the values of the original, see Fig. 1). For instance, when we qualify the price of a given item in

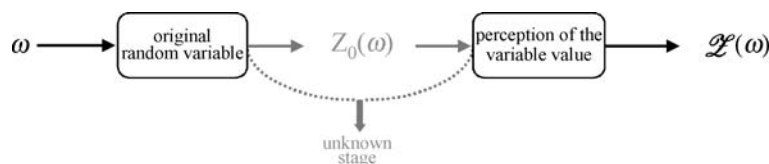
a specific store, we can perceive/label it as being ‘cheap’, but there is an existing price (although assumed to be unknown for the person receiving and processing data information).

Puri and Ralescu [31] introduced the concept of the also called fuzzy random variable as a generalization of the concept of random set or set-valued random element (and hence, as a generalization also of the concept of random variable). According to this definition a fuzzy random variable  $Z$  may be considered as a random element associating with each experimental outcome a value which is intrinsically fuzzy. Below, we recall Puri and Ralescu’s definition

**Definition 3 (Puri and Ralescu [31])** Given a probability space  $(\Omega, \mathcal{A}, P)$ , a mapping  $X : \Omega \rightarrow \mathcal{F}_c(\mathbb{R})$  is said to be a *fuzzy random variable* (also referred to as *random fuzzy set*) if for each  $\alpha \in [0, 1]$  the set-valued mapping  $X_\alpha : \Omega \rightarrow \mathcal{K}_c(\mathbb{R})$ , where  $\mathcal{K}_c(\mathbb{R})$  is the class of the nonempty compact intervals and  $X_\alpha(\omega) = (X(\omega))_\alpha$  for all  $\omega \in \Omega$ , is a compact convex random sets (that is, a Borel-measurable mapping with respect to the Borel  $\sigma$ -field generated by the topology associated with the Hausdorff metric on  $\mathcal{K}_c(\mathbb{R})$ ).

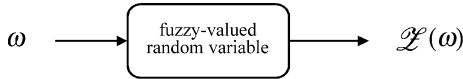
**Remark 1** The notion of fuzzy random variable has been in fact introduced in a more general way by considering as the codomain the space of fuzzy sets of the  $p$ -dimensional Euclidean space, or even more general Banach spaces, whose  $\alpha$ -levels are nonempty compact subsets of this space. In case one constrains to  $p = 1$  and fuzzy sets being convex, then one gets the last definition.

**Remark 2** Although motivation to introduce fuzzy random variables was different in the approaches by Kwakernaak/Kruse and Meyer and by Puri and Ralescu, one can prove that the notion in Definition 3 implies the one in Definition 2. As a consequence, probabilistic ideas and results for the notion in Definition 3 (or for the more general one in Remark 1) apply to the notion in Definition 2, and the same happens for statistical developments. However, many probabilistic conclusions, and most of the statistical procedures for Definition 2 are based on the assumption



Statistics with Imprecise Data, Figure 1

Fuzzy random variables in Kwakernaak/Kruse and Meyer’s sense



Statistics with Imprecise Data, Figure 2

Fuzzy random variables in Puri and Ralescu's sense

of having an unknown but existing original, and considering Zadeh's extension principle, so that these conclusions and procedures are not usually applicable to deal with data coming from fuzzy random variables in Definition 3.

**Remark 3** The concept of fuzzy random variable in Definition 3 can be alternatively formalized as a Borel-measurable mapping with respect to the Borel  $\sigma$ -field generated by the topology associated with some metrics on the space  $\mathcal{F}_c(\mathbb{R})$ , among them an operational one we will later refer to. Borel-measurability allows us to guarantee that notions like those of the induced distribution by a fuzzy random variable, independence of fuzzy random variables, identically distributed fuzzy random variables, and so on, can be immediately formalized in the probabilistic setting.

Values of a fuzzy random variable in Definition 3 have been conceived to model existing fuzzy values (see Fig. 2). For instance, when we classify a client of a bank in accordance with the degree of aversion to investment as having a 'rather high' degree, there is no underlying real-valued degree, but the classification itself is essentially imprecise.

### Statistical Analysis of Fuzzy Data Corresponding to Fuzzy Perceptions of Existing Real-Valued Data

#### Fuzzy Estimation

When imprecise statistical data correspond to fuzzy perceptions of unobserved/unknown precise (i. e. crisp) statistical data we can treat them as observed values of fuzzy random variables in the sense of Kwakernaak/Kruse and Meyer. In such a case we can analyze imprecise data in terms of probability distributions of their unobserved originals in a similar way as precise statistical data are analyzed using methods of traditional mathematical statistics. The only difference stems from the fact that having imprecise input information in the form of fuzzy data we cannot precisely evaluate the characteristics of the underlying probability distribution. Therefore, instead of finding precise values of the estimators of the 'parameters' describing the underlying probability distribution (the one of the original), it seems more coherent finding their imprecise fuzzy perceptions.

Assume that we observe a fuzzy random sample  $X_1, \dots, X_n$  which is viewed as a fuzzy perception of an

unobserved random sample  $X_1, \dots, X_n$ . Let  $F(x; \theta)$  be the cumulative probability function of the original random variable  $X$  characterized by a crisp parameter  $\theta \in \Theta$ . Suppose now that an estimator of  $\theta$  which is given by a statistic  $\hat{\theta} = \phi(X_1, \dots, X_n)$  is considered. By using Zadeh's extension principle we can consider as a *fuzzy estimator* of  $\theta$  based on  $X_1, \dots, X_n$  the one associating with each fuzzy sample information  $(\tilde{x}_1, \dots, \tilde{x}_n)$  the *fuzzy estimate*  $\phi(\tilde{x}_1, \dots, \tilde{x}_n)$  given by

$$\mu_{\phi(\tilde{x}_1, \dots, \tilde{x}_n)}(t) = \begin{cases} \sup_{(x_1, \dots, x_n) | t = \phi(x_1, \dots, x_n)} \min\{\mu_{\tilde{x}_1}(x_1), \dots, \mu_{\tilde{x}_n}(x_n)\} & \text{if } t \in \text{Im}(\phi(X_1, \dots, X_n)) \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

Alternatively,  $\phi(\tilde{x}_1, \dots, \tilde{x}_n)$  can be viewed as a fuzzy estimate of the induced fuzzy parameter  $\vartheta = \theta(X)$  with  $\theta(X)(t) = \sup_{X \in \mathcal{F}(\Omega, \mathcal{A}, P) | \theta(X)=t} \inf_{\omega \in \Omega} \mu_{X(\omega)}(X(\omega))$  and  $\mathcal{F}(\Omega, \mathcal{A}, P)$  being the class of all possible originals of  $X$ . In many practical cases, when  $\phi(x_1, \dots, x_n)$  has a simple form, the calculation of (7) is straightforward. For example, when  $\phi(x_1, \dots, x_n) = \text{sample mean} = (x_1 + \dots + x_n)/n$  the  $\alpha$ -levels of the estimate of its expected value  $\theta$  are expressed as

$$(\phi(\tilde{x}_1, \dots, \tilde{x}_n))_\alpha = \left[ \frac{1}{n} \sum_{i=1}^n (\tilde{x}_i)_L(\alpha), \frac{1}{n} \sum_{i=1}^n (\tilde{x}_i)_U(\alpha) \right] \quad (8)$$

Analogously, whenever the estimator of  $\theta$  is given by a continuous function or an injective function, the  $\alpha$ -levels becomes also rather simple. In other cases, the limits of the  $\alpha$ -levels should often be found by solving nonlinear mathematical programming problems defined by (7).

A similar approach may be applied when we try to construct fuzzy versions of the confidence intervals of the unknown parameter  $\theta$ . Let us assume that we are able to find confidence intervals of  $\theta$  using precise (crisp) statistical data. For example, let  $[\underline{\pi}_l, +\infty)$ , where  $\underline{\pi}_l = \underline{\pi}_l(X_1, \dots, X_n; \delta)$ , be the one-sided confidence interval for the parameter  $\theta$  on the confidence level  $1 - \delta$ . Kruse and Meyer [20] have shown that when we replace  $\underline{\pi}_l$  with the lower limits of the  $\alpha$ -cuts of its fuzzy version we obtain a proper confidence interval for the fuzzy perception of  $\theta$ . These lower limits can be found for different values of  $\alpha \in (0, 1]$  from the formula [15]

$$\begin{aligned} \underline{\pi}_\alpha^L &= \underline{\pi}_\alpha^L(X_1, \dots, X_n; \delta) \\ &= \inf \{t \in \mathbb{R} : \forall i \in \{1, \dots, n\} \exists x_i \in X_{i,\alpha} \\ &\quad \text{such that } \underline{\pi}_l(x_1, \dots, x_n; \delta) \leq t\} \end{aligned} \quad (9)$$

where  $X_{i,\alpha}$ ,  $i = 1, \dots, n$  are respective  $\alpha$ -cuts of the observed fuzzy data.

In a similar way we can define a fuzzy equivalent of the one-sided confidence interval  $(-\infty, \bar{\pi}_u]$ :

$$\begin{aligned} \bar{\Pi}_\alpha^U &= \bar{\Pi}_\alpha^U(X_1, \dots, X_n; \delta) \\ &= \sup \{t \in \mathbb{R} : \forall i \in \{1, \dots, n\} \exists z_i \in X_{i,\alpha} \\ &\quad \text{such that } \bar{\pi}_u(z_1, \dots, z_n; \delta) \geq t\} \end{aligned} \quad (10)$$

where  $\bar{\pi}_u(z_1, \dots, z_n; \delta) = \underline{\pi}_l(z_1, \dots, z_n; 1 - \delta)$ . Moreover, exactly the same approach can be applied when we look for two-sided confidence intervals.

Summing up this section we can say that when imprecise observations may be treated as fuzzy perceptions of precise but unobserved realizations of ordinary random variables the problem of point and interval estimation of the unknown parameters of the underlying probability distribution can be reduced to finding fuzzy versions of the formulae known from traditional mathematical statistics.

### Fuzzy Statistical Tests

Testing statistical hypotheses is the second main branch of mathematical statistics. Tests of statistical hypotheses have to be applied if we want to make decisions based on the analysis of random data. When our decisions depend on the values of the parameters of probability distributions that describe observed statistical data we use parametric statistical methods. In such a case we test statistical hypotheses about the values of the parameters of probability distributions utilizing a well known equivalence between the set of values of the considered probability distribution parameter for which the null hypothesis is accepted and a certain confidence interval for this parameter. Kruse and Meyer [20] have shown that the same equivalence exists in the case of statistical tests with fuzzy data.

Let  $X_1, \dots, X_n$  denote a fuzzy sample, i.e. a fuzzy perception of the usual random sample  $X_1, \dots, X_n$ , from the population with the distribution  $P_\theta$ . Let  $\delta$  be a given number from the interval  $(0, 1)$ . Grzegorzewski ([15]) proposed the following definition of the *fuzzy test* for vague data:

**Definition 4** A function  $\varphi : (\mathcal{F}_c(\mathbb{R}))^n \rightarrow \mathcal{F}(\{0, 1\})$  is called a fuzzy test for the hypothesis  $H$ , at the significance level  $\delta$ , if

$$\sup_{\alpha \in [0, 1]} P \{ \omega \in \Omega : \varphi_\alpha(X_1(\omega), \dots, X_n(\omega)) \subseteq \{1\} | H \} \leq \delta, \quad (11)$$

where  $\varphi_\alpha$  is the  $\alpha$ -level set ( $\alpha$ -cut) of  $\varphi(X_1, \dots, X_n)$ .

The fuzzy test defined above can be regarded as a family of classical tests  $\{\varphi_\alpha : \alpha \in (0, 1]\}$  for which the significance level is given as the upper bound of type I error for the whole family  $\{\varphi_\alpha : \alpha \in (0, 1]\}$ .

In order to give an example of a fuzzy statistical test let us consider a following simple null hypothesis:  $H : \theta = \theta_0$ , against the composite two-sided alternative:  $K : \theta \neq \theta_0$ . Suppose we know a two-sided symmetrical confidence interval  $[\pi_1, \pi_2]$  for  $\theta$ , on a confidence level  $1 - \delta$ , where  $\pi_1 = \pi_1(X_1, \dots, X_n; \delta/2)$  and  $\pi_2 = \pi_2(X_1, \dots, X_n; \delta/2)$  are the limits of the ordinary two-sided confidence interval. The fuzzy equivalent of this confidence interval can be calculated using the  $\alpha$ -cuts  $\Pi_\alpha = [\Pi_\alpha^L, \Pi_\alpha^U]$  for all  $\alpha \in (0, 1]$ , where the limits of these  $\alpha$ -cuts can be computed from Eq. (9)-(10) by replacing  $\delta$  with  $\delta/2$ . The fuzzy two-sided statistical test for  $H : \theta = \theta_0$ , against  $K : \theta \neq \theta_0$ , on the significance level  $\delta$ , has been defined by Grzegorzewski [15] as a function  $\varphi : (\mathcal{F}_c(\mathbb{R}))^n \rightarrow \mathcal{F}(\{0, 1\})$  with following  $\alpha$ -cuts

$$\varphi_\alpha(X_1, \dots, X_n) = \begin{cases} \{0\} & \text{if } \theta_0 \in (\Pi_\alpha \setminus (\neg \Pi)_\alpha), \\ \{1\} & \text{if } \theta_0 \in ((\neg \Pi)_\alpha \setminus \Pi_\alpha), \\ \{0, 1\} & \text{if } \theta_0 \in (\Pi_\alpha \cap (\neg \Pi)_\alpha), \\ \emptyset & \text{if } \theta_0 \notin (\Pi_\alpha \cup (\neg \Pi)_\alpha), \end{cases} \quad (12)$$

Similarly we may obtain fuzzy tests for one-sided hypotheses using the one-to-one correspondence between the acceptance regions of the tests designated for testing these hypotheses on the significance level  $\delta$  and one-sided confidence intervals for the parameter  $\theta$  on the confidence level  $1 - \delta$ .

Grzegorzewski [15] has shown that the membership function of the fuzzy test for the hypothesis  $H$  against  $K$  is given by

$$\begin{aligned} \mu_\varphi(t) &= \mu_\Pi(\theta_0)I_{\{0\}}(t) + \mu_{\neg \Pi}(\theta_0)I_{\{1\}}(t) \\ &= \mu_\Pi(\theta_0)I_{\{0\}}(t) + (1 - \mu_\Pi(\theta_0))I_{\{1\}}(t), \quad t \in \{0, 1\}, \end{aligned} \quad (13)$$

where  $\Pi$  is a fuzzy acceptance region depending on the considered hypotheses. Thus, the fuzzy fuzzy test defined by Eq. (12), contrary to the classical crisp test, does not lead to the binary decision – to accept or to reject the null hypothesis – but to a fuzzy decision. One may get  $\mu_\varphi(0) = 1, \mu_\varphi(1) = 0$  which indicates that we should accept  $H$ , or  $\mu_\varphi(0) = 0, \mu_\varphi(1) = 1$  which means the rejection of  $H$ . However, one may also get  $\mu_\varphi(0) = \mu_0, \mu_\varphi(1) = 1 - \mu_0$ , where  $\mu_0 \in (0, 1)$ , which can be interpreted as a degree of conviction that we should accept ( $\mu_0$ ) or reject ( $1 - \mu_0$ ) the hypothesis  $H$ . Thus, in



situation when  $\mu_0$  is neither 0 nor 1, a user must decide using other criteria whether to reject or to accept the considered hypothesis. There exist several approaches that are suitable for solving this problem. One of these approaches which is formulated in the language of the possibility theory has been proposed by Hryniewicz [16] who used the results of Dubois et al. [9] who proposed to use statistical confidence intervals of parameters of probability distributions for the construction of possibility distributions of these parameters. According to their approach, the family of two-sided confidence intervals

$$[\pi_L(x_1, \dots, x_n; 1 - \delta/2), \pi_U(x_1, \dots, x_n; 1 - \delta/2)], \quad \delta \in (0, 1) \quad (14)$$

forms the *possibility distribution*  $\tilde{\vartheta}$  of the estimated value of the unknown parameter  $\vartheta$ . In a similar way it is possible to construct one-sided possibility distributions based on one-sided nested confidence intervals. Hryniewicz [16] proposed to compare this possibility distribution with a hypothetical value of the tested parameter. For this purpose he proposed to use the necessity of strict dominance measure introduced by Dubois and Prade [7] for measuring the necessity of the strict dominance relation  $\tilde{A} \succ \tilde{B}$ , where  $\tilde{A}$  and  $\tilde{B}$  are fuzzy sets. This measure, called the *Necessity of Strict Dominance index (NSD)*, is defined as

$$\begin{aligned} NSD &= \text{Ness}(\tilde{A} \succ \tilde{B}) \\ &= 1 - \sup_{x, y; x \leq y} \min \{ \mu_A(x), \mu_B(y) \}. \end{aligned} \quad (15)$$

Hryniewicz [16] has shown that in the classical case of precise statistical data and precisely defined statistical hypotheses the value of the *NSD* index is equal to the *p*-value of the test.

In case of fuzzy data the confidence intervals used for the construction of the possibility distribution of the estimated parameter  $\theta$  can be replaced by their fuzzy equivalents presented in the previous sections of this article. In his paper Hryniewicz [16] assumes that the value of the significance level of the corresponding statistical test  $\delta$  is equal to the possibility degree  $\alpha$  that defines the respective  $\alpha$ -cut of the possibility distribution of  $\tilde{\theta}$ . He also assumes that in the possibilistic analysis of statistical tests on the significance level  $\delta$  we should take into account only those possible values of the fuzzy sample whose possibility is not smaller than  $\delta$ . Thus, the  $\alpha$ -cuts of the membership function  $\mu_F(\theta)$  denoted by  $[\mu_{F,L}^{(\alpha)}, \mu_{F,U}^{(\alpha)}]$  are equivalent to the  $\alpha$ -cuts of the respective fuzzy confidence intervals on a confidence level  $1 - \alpha$ . Having the possibility distribution of the test statistic we can use Eq. (15) for the calculation of the degree on necessity that the considered sta-

tistical hypothesis has to be accepted. When we set a critical value for this characteristic we arrive at unequivocal (crisp) decisions. It is worthy to note that this approach has been generalized in [16] to the case of testing imprecisely defined hypothesis using fuzzy statistical data.

In the previous paragraphs we have presented fuzzy statistical tests when the class the underlying probability distribution belongs to is known. Verification of this assumption when the available statistical data are imprecise may be very difficult indeed. Therefore, it would be advisable to use fuzzy equivalents of non-parametric (distribution-free) statistical methods. Unfortunately, there exist only few papers devoted to such fuzzy tests. The most interesting result has been obtained by Denœux et al. [4] who proposed a general methodology for the construction of fuzzy tests based on rank statistics.

Statistical analysis of fuzzy random data can be also done in the Bayesian framework. First results presenting the Bayesian decision analysis for imprecise data were given in papers by Casals et al. [3] and Gil [11]. Other approaches have been proposed by such authors as Viertl [34], Frühwirth-Schnatter [10], and Taheri and Behboodian [32]. Comprehensive Bayesian model comprising fuzzy data, fuzzy hypotheses, and fuzzy utility function has been proposed in the paper by Hryniewicz [17].

### Statistical Analysis of Existing Fuzzy-Valued Data

When imprecise statistical data correspond to intrinsic fuzzy-valued data we can treat them as observed values of fuzzy random variables in the sense of Puri and Ralescu. In the literature these data are often treated as categorical/ordinal/interval-valued ones. It should be emphasized that the model given by fuzzy random variables allows us to describe and handle these data in a more expressive scale and way (in contrast to just ranking or stating simply the interval support of the values). Thus, many statistical developments for real-valued data are based on distances/deviations between values rather than on the diversity of these values. The use of the fuzzy scale allows to consider metrics with a meaning similar to that for the real-valued case (i. e., distinguishing not only the ranks of variable values w.r.t. a certain criterion, but a physical distance between them).

The distance we will consider here is the one stated by Bertoluzza et al. [2], so that if  $A, B \in \mathcal{F}_c(\mathbb{R})$

$$\begin{aligned} D_W^\varphi(A, B) &= \\ &= \sqrt{\int_{[0,1]} \left[ \int_{[0,1]} [f_A(\alpha, \lambda) - f_B(\alpha, \lambda)]^2 dW(\lambda) \right] d\varphi(\alpha)} \end{aligned} \quad (16)$$

with  $f_A(\alpha, \lambda) = \lambda \cdot A_U(\alpha) + (1 - \lambda) \cdot A_L(\alpha)$ , where

- $W$  and  $\varphi$  are normalized weighted measures on  $[0, 1]$  formalized as probability measures on  $([0, 1], \mathcal{B}_{[0,1]})$ ,
- $W$  is associated with a non-degenerate distribution,
- $\varphi$  is associated with a strictly increasing distribution function on  $[0, 1]$ .

*Remark* It should be remarked on  $D_W^\varphi$  that

- $W$  and  $\varphi$  have no stochastic meaning.
- To consider  $W$  is equivalents to consider a measure weighting points 0, 1 and a certain  $t_0(W) \in (0, 1)$ ; in case  $W$  is symmetric, then  $t_0(W) = .5$ .
- For each  $\alpha$ , the choice of  $W$  allows us to weight
  - the effect of the distance between the “widths” of the  $\alpha$ -levels (i. e., effect of the “shape” difference),
  - in comparison with the effect of the distance between their  $t_0(W)$ -points (i. e., effect of the “location” difference).
- The choice of  $\varphi$  allows us to weight the influence of each level (i. e., degree of “imprecision”, “consensus”, “subjectivity”, ...).
- $D_W^\varphi$  is a versatile and operational in statistical developments with fuzzy numbers, and it behaves especially well when we consider least-squares approaches.

Since the concept of fuzzy random variable in Puri and Ralescu’s sense has been properly stated in a probabilistic context as a random element (i. e., as a Borel-measurable function), concepts like independent and identically distributed fuzzy random variables make immediate sense. Furthermore, all the main ideas, aims and concepts, and several developments can be immediately considered to deal with fuzzy data when coming from fuzzy-valued Borel-measurable mappings. In this respect, notions like either unbiasedness or consistency of a “point” (fuzzy- or real- valued) estimator of a (fuzzy- or real- valued) ‘parameter’ associated with the distribution of the fuzzy random variable, or the  $p$ -value and power of a test concerning such a ‘parameter’, make the same sense as in the classical case.

Several developments have been made in connection with both estimation and testing of fuzzy- and real-valued parameters associated with the distribution of a fuzzy random variable in Puri and Ralescu’s sense. We are now just recalling a few results concerning the “point” fuzzy estimation and testing of the population mean of a fuzzy random variable, which is formalized as follows:

**Definition 5 (Puri & Ralescu, [31])** Given a probability space  $(\Omega, \mathcal{A}, P)$  and an associated fuzzy random variable

$\mathcal{X} : \Omega \rightarrow \mathcal{F}_c(\mathbb{R})$  such that  $\max \{ |\inf \mathcal{X}_0|, |\sup \mathcal{X}_0| \}$  is integrable, then, the **fuzzy expected value** (or **fuzzy mean**) of  $\mathcal{X}$  is the fuzzy number  $\tilde{E}(\mathcal{X}|P) \in \mathcal{F}_c(\mathbb{R})$  such that for all  $\alpha \in [0, 1]$

$$\begin{aligned} & (\tilde{E}(\mathcal{X}|P))_\alpha \\ &= \text{Aumann integral of } \mathcal{X}_\alpha \\ &= \{E(\mathcal{X}|P) | \mathcal{X} : \Omega \rightarrow \mathbb{R}, \\ & \quad \mathcal{X} \in L^1(\Omega, \mathcal{A}, P), \mathcal{X} \in \mathcal{X}_\alpha \text{ a.s. } [P]\} \\ &= [E(\inf \mathcal{X}_\alpha | P), E(\sup \mathcal{X}_\alpha | P)] . \end{aligned}$$

*Remark 5* The fuzzy mean satisfies that

- If  $\mathcal{X}(\Omega) = \{\tilde{x}_1, \dots, \tilde{x}_m, \dots\} \subset \mathcal{F}_c(\mathbb{R})$ , then, it is coherent with fuzzy arithmetic, that is,

$$\begin{aligned} \tilde{E}(\mathcal{X}|P) &= P(\{\omega \in \Omega | \mathcal{X}(\omega) = \tilde{x}_1\}) \cdot \tilde{x}_1 + \dots \\ &+ P(\{\omega \in \Omega | \mathcal{X}(\omega) = \tilde{x}_m\}) \cdot \tilde{x}_m + \dots \end{aligned}$$

- Strong Laws of Large Numbers are satisfied for different metrics (like  $D_W^\varphi$ , and stronger ones), which also corroborates the suitability of the defined fuzzy mean as the stochastic limit of the sample ones.
- $\tilde{E}(\mathcal{X}|P)$  is the “Fréchet expectation” of  $\mathcal{X}$  w.r.t.  $D_W^\varphi$ , i. e., for all  $A \in \mathcal{F}_c(\mathbb{R})$ :

$$E\left([D_W^\varphi(\mathcal{X}, \tilde{E}(\mathcal{X}|P))]^2 | P\right) \leq E\left([D_W^\varphi(\mathcal{X}, A)]^2 | P\right) .$$

The interpretation of the fuzzy mean becomes more clear if we notice that for interval-valued random variables (i. e., random intervals) the expected value is given in a form of an interval with the lower limit equal to the expected value of the lower limits of observed random variables, and the upper limit equal to the expected value of the upper limits of observed random variables. Therefore, the expected value of the fuzzy random variable can be given as a nested set of such intervals represented by respective  $\alpha$ -cuts, as it is formally described in the definition.

The definition of other characteristics describing fuzzy random variables may be much more complicated. For example, the definition of the variance requires the introduction of the “Fréchet expectation” (defined above), as it was proposed by Körner and Näther [24].

### Fuzzy Estimation

Assume that we observe a fuzzy simple random sample  $\mathcal{X}_1, \dots, \mathcal{X}_n$  which is viewed now as an  $n$ -tuple of  $n$  independent fuzzy random variables which are identically distributed as  $\mathcal{X}$ . The associated **fuzzy sample mean** is the

statistic given by

$$\bar{X}_n = \frac{1}{n} \cdot [X_1 + \dots + X_n] .$$

Then, one can prove that (see Lubiano ([25]))

**Theorem 1** *The fuzzy sample mean satisfies that*

- i)  $\bar{X}_n[\cdot]$  in an “unbiased fuzzy-valued estimator” of  $\tilde{E}(X|P)$  (in the sense of the fuzzy expected value defined by Puri & Ralescu). For most of the metrics we can consider, it is also a ‘strongly consistent’ fuzzy-valued estimator of  $\tilde{E}(X|P)$ .
- ii) One can quantify the mean squared-type error in the fuzzy estimation by considering the real-valued expected value  $E\left([D_W^\varphi(\bar{X}_n, \tilde{E}(X|P))]^2\right)$ .

Actually, Lubiano et al. [26,27] proposed several developments in connection with the estimation and testing about the  $D_W^\varphi$ -mean squared error associated with the estimation of the population fuzzy means by means of the sample one.

### Statistical Tests

One of the problems which has received a deep attention in connection with testing from fuzzy random variables in Puri and Ralescu’s sense is that concerning two-sided tests about the mean of a fuzzy random variable (one-sample case). In order to define this problem one can decide which metrics could be used for measuring the distance between the hypothetical and observed fuzzy values of considered characteristics of the fuzzy random variables. In the case of the metrics defined by Eq. (16), the problem can be formalized as follows:

Given  $n$  independent observations from  $X$ ,  $X_1, \dots, X_n$ , we wish to test the null hypothesis  $H_0: \tilde{E}(X|P) = A \in \mathcal{F}_c(\mathbb{R})$ , which can be equivalently expressed as  $H_0: D_W^\varphi(\tilde{E}(X|P), A) = 0$ .

Despite simple formulation of the testing problem the construction of statistical tests for the verification of hypotheses related to fuzzy mean values is not that simple. For example, an exact test has been developed in Montenegro et al. [28] for so called “normal” fuzzy random variables (in Puri and Ralescu’s sense, [30]). Although the method is exact and easy-to-apply, the assumption of  $X$  being fuzzy “normal” ( $X = V + \mathcal{N}(0, 1)$ , with  $V \in \mathcal{F}_c(\mathbb{R})$ ) is quite restrictive and often unrealistic, as it means that *all* observed fuzzy values are described by the same membership function shape with maybe different location.

In a more general case asymptotic tests based on Large Sample Theory have been developed for the same purpose. However, they are usually hardly applicable (except for some simple special cases) in practice. For example, the asymptotic distribution of the statistic proposed by Körner [23] involves unknown parameters such as correlations between random variables describing membership functions of observed fuzzy variables. In the paper by Montenegro et al. [28] devoted to the problem of testing the equality of two fuzzy mean values, it is assumed that  $X$  takes on a finite number of values, and large sample sizes would be required anyway in order to estimate unknown parameters of the model.

Simulation studies have been considered to analyze the extent and applicability of the asymptotic test by Montenegro et al. [28]. These studies have confirmed that in estimating the unknown parameters (the eigenvalues of a certain correlation matrix) entails a substantial loss of precision w.r.t. the nominal significance level. Based on this empirical conclusion, the use of  $D_W^\varphi$  and the Generalized Bootstrapped CLT (Giné and Zinn, [13]) allow us to consider bootstrap techniques in this context.

The bootstrap technique consists in taking random samples (i. e. re-sampling) from the original one and calculating the value of the considered statistic. By repeating this procedure many times we obtain the sample distribution of the test statistic which may be thus used for testing purposes. In case of fuzzy random data we get the following method proposed by González-Rodríguez et al. [14]:

**Theorem 2** *Given a fuzzy random variable  $X: \Omega \rightarrow \mathcal{F}_c(\mathbb{R})$  associated with the probability space  $(\Omega, \mathcal{A}, P)$  and such that*

- $\max\{(\inf X_0)^2, (\sup X_0)^2\}$  is integrable ,
- $X_1, \dots, X_n$  are i.i.d. as  $X$  ,
- $X_1^*, \dots, X_n^*$  is a bootstrap sample from  $X_1, \dots, X_n$  ,

*then, to test  $H_0$  at the nominal significance level  $\alpha \in [0, 1]$ ,  $H_0$  should be rejected whenever*

$$\frac{[D_W^\varphi(\bar{X}_n, U)]^2}{\hat{S}_n^2} > z_\alpha ,$$

*where  $z_\alpha$  is the  $100(1 - \alpha)$  fractile of the bootstrap distribution of*

$$T_n = [D_W^\varphi(\bar{X}_n^*, \bar{X}_n)]^2 / \hat{S}_n^{*2}$$

*with*

$$\bar{X}_n^* = \sum_{i=1}^n X_i^* / n, \quad \hat{S}_n^{*2} = \sum_{i=1}^n [D_W^\varphi(X_i^*, \bar{X}_n^*)]^2 / (n-1) .$$

Other bootstrap tests about means of fuzzy random variables which have been recently developed are the following ones:

- One-sided hypotheses tests in the one-sample case.
- Tests for the equality of means of two FRVs
  - for two independent samples
  - for two linked samples.
- Tests for the equality of means of  $J$  FRVs (ANOVA):
  - for  $J$  independent samples.

From comparative extensive simulation studies that have been performed recently at the University of Oviedo and the European Center for Soft Computing we can draw the following practical conclusions:

- For small/medium samples, the bootstrap method performs and behaves much better than the asymptotic one.
- For large sample sizes (over 300), the improvement is not that remarkable, but the bootstrap approach still provides the best approximation to the nominal significance level.

Taking into account that fuzzy statistical test for testing hypotheses about other characteristics of fuzzy random variables are even more complicated than the procedures designed for testing the hypotheses about mean values we can claim that the bootstrap technique is probably the most promising for dealing with random fuzzy observations.

### Future Directions

Statistical analysis of imprecise data is still a developing area of science. Future directions of its development are tightly connected with the development of methods that may be used for the description of uncertainties of different types. It has been pointed out by P. Walley (see, e.g. [36] for a good overview) that traditional probability is not sufficient for good description of different types of uncertainty. Different mathematical models, such as e.g. Dempster–Shafer belief functions, possibility distributions, lower and upper probabilities, lower and upper previsions, and many others, have been proposed for this purpose. However, for the most general models describing uncertainty appropriate statistical methods have not been proposed yet. Therefore, statistical methods for handling very general imprecise data have to be developed in the future.

Another, but much more specific, future direction for the development of statistical analysis of imprecise data is

related to the analysis of intrinsically fuzzy data. In contrast to the situation when fuzzy observations may be considered as fuzzy perceptions of real-valued observations, many notions known from traditional statistics are still waiting for their widely accepted definitions, and statistical methods of analysis.

The most challenging future direction is related to Zadeh's paradigm of "computing with words". First of all, we need operational methods for the representation of linguistic concepts which could be useful is statistical analysis of imprecisely reported (with words!) statistical data. Moreover, we also need methods for convincing presentation of the results of computations to users who have only limited knowledge of mathematics and statistics.

### Bibliography

#### Primary Literature

1. Bandemer H, Näther W (1992) Fuzzy Data Analysis. Kluwer, Dordrecht
2. Bertoluzza C, Corral N, Salas A (1995) On a new class of distances between fuzzy numbers, *Mathware and Soft Computing*, vol 2. Departament of Computer Science and Artificial Intelligence of the University of Granada and Secció de Matemàtiques i Informàtica of the Universitat Politècnica de Catalunya, Granada, Barcelona, pp 71–84. <http://docto-si.ugr.es/Mathware/ENG/mathware.html>
3. Casals R, Gil MA, Gil P (1986) The fuzzy decision problem. In: An approach to the problem of testing statistical hypotheses with fuzzy information. *Eu J Oper Res* 27:371–382
4. Denœux T, Masson M-H, Hébert PA (2005) Nonparametric rank-based statistics and significance tests for fuzzy data. *Fuzzy Sets Syst* 153:1–28
5. Dubois D, Prade H (1978) Operations on Fuzzy Numbers. *Int J Syst Sci* 9:613–626
6. Dubois D, Prade H (1980) Fuzzy Sets and Systems. In: *Theory and Applications*. Academic Press, New York
7. Dubois D, Prade H (1983) Ranking fuzzy numbers in the setting of possibility theory. *Inf Sci* 30:184–244
8. Dubois D, Prade H (1988) Possibility Theory. Plenum Press, New York
9. Dubois D, Foulloy L, Mauris G, Prade H (2002) Probability-possibility transformations, triangular fuzzy-sets and probabilistic inequalities, *Proc of the Ninth International Conference IPMU*, Annecy, pp 1077–1083
10. Frühwirth-Schatter S (1993) Fuzzy Bayesian inference. *Fuzzy Sets Syst* 60:41–58
11. Gil MÁ, Kacprzyk J, Fedrizzi M (eds) (1988) Probabilistic-Possibilistic approach to some Statistical Problems with Fuzzy Experimental Observations. In: *Combining Fuzzy Imprecision with Probabilistic Uncertainty in Decision Making*. Springer, Berlin, pp 286–306
12. Gil MÁ, López-Díaz M, Ralescu DA (2006) Overview on the development of fuzzy random variables. *Fuzzy Sets Syst* 157:2546–2557
13. Giné E, Zinn J (1990) Bootstrapping general empirical measures. *Ann Prob* 18:851–869

14. González-Rodríguez G, Montenegro M, Colubi A, Gil MA (2006) Bootstrap techniques and fuzzy random variables. Synergy in hypothesis testing with fuzzy data. *Fuzzy Sets Syst* 157: 2608–2613
15. Grzegorzewski P (2000) Testing statistical hypotheses with vague data. *Fuzzy Sets Syst* 112:501–510
16. Hryniewicz O (2006) Possibilistic decisions and fuzzy statistical tests. *Fuzzy Sets Syst* 157:2665–2673
17. Hryniewicz O, Grzegorzewski P, Gil MÁ (eds) (2002) Possibilistic Approach to the Bayes Statistical Decisions. In: *Soft Methods in Probability, Statistics and Data Analysis*. Physica, Heidelberg, pp 207–218
18. ISO/IEC (1995) Guide to the expression of uncertainty in measurement (GUM). ISO/IEC, Geneva
19. Kruse R (1982) The strong law of large numbers for fuzzy random variables. *Inf Sci* 28:233–241
20. Kruse R, Meyer KD (1987) *Statistics with Vague Data*. D. Riedel Publishing Company, Dordrecht
21. Kwakernaak H (1978) Fuzzy Random Variables. Definitions and Theorems. *Inf Sci* 15:1–15
22. Kwakernaak H (1979) Fuzzy Random Variables. Algorithms and Examples for the Discrete Case. *Inf Sci* 17:253–278
23. Körner R (2000) An asymptotic  $\alpha$ -test for the expectation of random fuzzy variables. *J Stat Plann Inference* 83:331–346
24. Körner R, Näther W (2002) On the variance of random fuzzy variables. In: Bertoluzza C, Gil MÁ, Ralescu DA (eds) *Statistical Modeling, Analysis and Management of Fuzzy Data*. Physica, Heidelberg, pp 25–42
25. Lubiano MA (1999) *Medidas de variación de elementos aleatorios*. Ph D Thesis, University of Oviedo
26. Lubiano MA, Gil MA (1999) Estimating the expected value of fuzzy random variables in random samplings from finite populations. *Stat Pap* 40:277–295
27. Lubiano MA, Gil MA, López-Díaz M, López-García MT (2000) The  $\vec{\lambda}$ -mean squared dispersion associated with a fuzzy random variable. *Fuzzy Sets Syst* 111:307–317
28. Montenegro M, Colubi A, Casals MR, Gil MA (2004) Asymptotic and Bootstrap techniques for testing the expected value of a fuzzy random variable. *Metrika* 59:31–49
29. Nguyen HT (1978) A note on the extension principle for fuzzy sets. *J Math Anal Appl* 64:369–380
30. Puri ML, Ralescu DA (1985) The concept of normality for fuzzy random variables. *Ann Prob* 13:1373–1379
31. Puri ML, Ralescu DA (1986) Fuzzy Random Variables. *J Math Anal Appl* 114:409–422
32. Taheri SM, Behboodian J (2001) A Bayesian approach to fuzzy hypotheses testing. *Fuzzy Sets Syst* 123:39–48
33. Terán P (2007) Probabilistic foundations for measurement modelling with fuzzy random variables. *Fuzzy Sets Syst* 158:973–986
34. Viertl R (ed) (1987) Is it necessary to develop a fuzzy Bayesian inference. In: *Probability and Bayesian Statistics*. Plenum, New York, pp 471–475
35. Viertl R (1996) *Statistical Methods for Non-Precise Data*. CRC Press, Boca-Raton
36. Walley P (1996) Measures of uncertainty in expert systems. *Artif Intell* 114:1–58
37. Zadeh LA (1956) Fuzzy sets. *Inf Control* 8:338–353
38. Zadeh LA (1975) The concept of a linguistic variable and its application to approximate reasoning. *Inf Sci* 8(1):199–249; 8(2):301–353; 9(3):43–80

## Books and Reviews

- Bertoluzza C, Gil MÁ, Ralescu DA (eds) (2002) *Statistical Modeling, Analysis and Management of Fuzzy Data*. Physica, Heidelberg
- Dubois D, Lubiano MA, Prade H, Gil MÁ, Grzegorzewski P, Hryniewicz O (eds) (2008) *Soft methods for handling variability and imprecision*. Springer, Berlin
- Gil MÁ, López-Díaz M, Ralescu DA (2006) Overview on the development of fuzzy random variables. *Fuzzy Sets Syst* 157: 2546–2557
- Grzegorzewski P, Hryniewicz O, Gil MÁ (eds) (2002) *Soft methods in probability*. In: *Statistics and data analysis*. Physica, Heidelberg
- Kruse R, Gebhardt J, Gil MÁ (1999) *Fuzzy Statistics*. In: Webster JG (ed) *Encyclopedia of Electrical and Electronics Engineering*. Wiley, New York
- Lawry J, Miranda E, Bugarin A, Li S, Gil MÁ, Grzegorzewski P, Hryniewicz O (eds) (2006) *Soft Methods for Integrated Uncertainty Modeling*. Springer, Berlin
- López-Díaz M, Gil MÁ, Grzegorzewski P, Hryniewicz O, Lawry J (eds) (2004) *Soft Methodology and Random Information Systems*. Springer, Berlin
- Taheri SM (2003) Trends in Fuzzy Statistics. *Austrian J Stat* 32:239–257

## Stellar Dynamics, *N*-body Methods for

JUNICHIRO MAKINO

Center for Computational Astrophysics, National Astronomical Observatory of Japan, Tokyo, Japan

## Article Outline

[Glossary](#)  
[Definition of the Subject](#)  
[Introduction](#)  
[Statistical Description](#)  
[Numerical Methods](#)  
[Future Directions](#)  
[Bibliography](#)

## Glossary

**Binary** Two stars which orbit around each other.

**Distribution function** A density function in the six-dimensional phase space which gives the distribution of stars in a stellar system.

**Fokker–Planck equation** Partial differential equation for the thermal evolution of the distribution function of a stellar system, expressed in the form of the advection-diffusion equation in the phase space.



**Monte-Carlo method** The method to numerically follow the thermal evolution of stellar systems using Monte-Carlo integration of Fokker–Planck equation for the distribution function.

**Relaxation** The process which leads the system to thermal equilibrium. In the case of a stellar system, the main mechanism of relaxation is the gravitational encounter between stars.

**Stellar systems** A system composed of a large number of stars, i. e., particles which interact through mutual gravity.

**Thermal stability** Stability of the system against the redistribution of the thermal energy.

## Definition of the Subject

Most of astronomical objects are in the first approximation the collection of point mass particles (stars or planets). We call such systems stellar systems, and stellar dynamics is the theoretical framework for the study of such systems. Since the equation of motion for stellar systems with  $N$  stars is analytically solvable only for the case of  $N = 2$ , numerical integration of the orbit of stars has been very important tool for the study of stellar systems.

## Introduction

Stellar Dynamics deals with the evolution of systems which consist of a large number of stars interacting with other stars through mutual gravitational force. Examples of such systems include star clusters, galaxies, clusters of galaxies. Star clusters consist of up to around 10 million stars. They are classified into two categories: open clusters and globular clusters. Open clusters are less massive and younger than globular clusters. Globular clusters in our galaxy are all very old (more than 10 Gyr) and fairly massive (more than  $10^5$  solar mass). Both types of clusters are compact, with the radius of the order of 10 parsec or less.

Galaxies are larger and more massive than star clusters. The total mass of our Galaxy is around  $10^{12}$  solar mass. Around 80% of the total mass is in dark matter, the existence of which we only indirectly know through the measurement of the motions of visible matter such as gas or stars.

In the first approximation, stellar systems such as star clusters and galaxies are just a collection of a large number of point masses. Therefore, one might expect that a statistical description is applicable. In other words, one might expect that the behavior of the system is described by the statistical mechanics. If that were the case, however, the universe would look very different from its present form. All galaxies would look similar, and rather boring, with-

out any structures such as spirals, bars, shells etc., which make the diversity of galaxies. In this article, we first review how the statistical description can be or cannot be applied to stellar systems, and then present an overview of our current understanding for the evolution of stellar systems. Then we discuss the numerical methods for  $N$ -body simulations. For a more complete coverage of the subject see textbooks [5,25].

## Statistical Description

### Thermal Stability and Gravothermal Catastrophe

There are two basic reasons why stellar systems cannot be described by statistical mechanics. The first one is that the thermal equilibrium state does not exist for stellar systems. An equilibrium state for a system of classical particles is given by the Maxwell-Boltzmann statistics. This means that the local velocity distribution is Maxwellian with a single temperature everywhere in the system. In other words, the velocity distribution function is given by

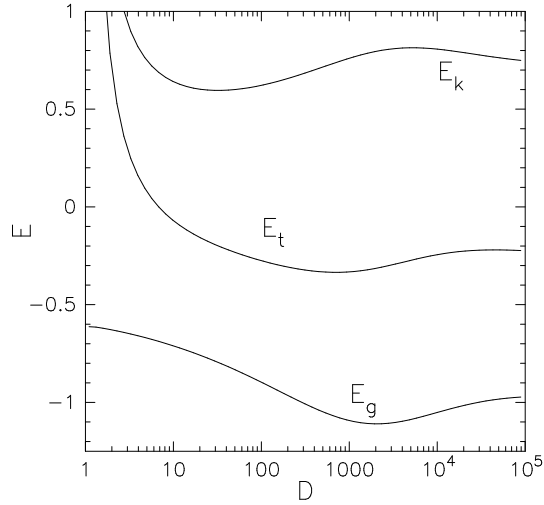
$$f_0(v) = \frac{n_f}{(2\pi\sigma^2)^{3/2}} \exp\left(\frac{-v^2/2}{\sigma^2}\right), \quad (1)$$

where  $n_f$  is the volume number density of particles and  $\sigma$  is the velocity dispersion.

However, with a stellar system, this Maxwellian distribution cannot be realized, simply because the stars with sufficiently high velocities will escape from the system. The depth of the gravitational potential of a stellar system is finite. Thus, if a star gained kinetic energy higher than that of the depth of the gravitational potential at its position, it would escape from the system.

As a thought experiment, we can think of a stellar system with a spherical adiabatic wall around it. The thermal equilibrium state of the stellar system can be described by that of an ideal gas, since the distribution function is Maxwellian for both cases. Therefore, even though we cannot apply the terms like pressure and temperature to stellar systems in general, we can apply these concepts to stellar systems in the thermal equilibrium, since the equilibrium distribution function is the same as that of an ideal gas with self gravity.

In this case, since the stars do not escape, the equilibrium state does exist. In the limit of very high temperature (large kinetic energy), the distribution of stars within the wall is uniform, with the number density of stars same everywhere. However, as we lower the temperature, the central density becomes higher due to the effect of self gravity. There must be a pressure gradient to support the gravity, and that means a density gradient should exist. However, this increase of density gradient means an increase of the



Stellar Dynamics, *N*-body Methods for, Figure 1

Energies as function of the ratio between the density at the center and that at the place of the wall,  $D$ . Three curves indicate the kinetic energy (top), total energy (middle) and gravitational energy (bottom), respectively

gravitational force, because more mass is concentrated to the center. Thus, if the increase of the gravitational force is larger than the increase of pressure, than the pressure gradient cannot support the system and there is no equilibrium state.

Figure 1 shows this phenomenon more clearly. Here, we plot the thermal energy ( $E_k$ ), gravitational energy ( $E_g$ ) and total energy ( $E_t = E_k + E_g$ ) as a function of the ratio of the density at the center and that of just inside the wall. Here, for simplicity, we use the system of units where the radius of the wall  $R$  and mass of the gas  $M$  are both unity and the gravitational constant  $G$  is also unity. The limit as  $D \rightarrow 1$  corresponds to infinite temperature. In this limit, the gravitational energy converges to  $-0.6$ , and the kinetic energy and total energy diverge. As we increase  $D$ , the kinetic energy decreases, gravitational energy increases, and the total energy decreases. However, at  $D = 709$ , the total energy  $E_t$  takes a minimum.

The fact that  $E_t$  takes a minimum means that this system has neutral stability against the perturbation which conserves the total energy, since an infinitesimal change of  $D$  in either direction does not change the total energy. In other words, for this system we can redistribute the thermal energy in such a way that after that perturbation is applied the system is still isothermal. For  $D < 709$ , when we change  $D$  the total energy also changes. In other words, if we redistribute the energy in any way, as far as we keep the total energy unchanged, the system cannot remain isothermal.

This change of behavior corresponds to a sign of the (second-order) variation of total entropy  $\delta^2 S$ . For  $D < 709$ , for any thermal perturbation,  $\delta^2 S < 0$ , and the second law of thermodynamics guarantees that the isothermal state is stable. However, for  $D = 709$ , there is one form of perturbation for which  $\delta^2 S = 0$ . For  $D > 709$ , there are one or more perturbations for which  $\delta^2 S > 0$ . In other words, the isothermal state is unstable for  $D > 709$ .

The eigenfunction which corresponds to  $\delta^2 S = 0$  has a single node, which means it either transfers the thermal energy from central area to the outer area, or vice versa. The system shows neutral stability against this heat transfer. Thus, even though the thermal energy is moved, the system remains isothermal.

In the case of  $D < 709$ , any perturbation in the distribution of the thermal energy results in the heat flux which cancels out the perturbation. In the case of  $D = 709$ , there is one mode of perturbation for which there will be no heat flux. Finally, in the case of  $D > 709$ , there is at least one mode of perturbation, for which the resulting heat flows in the direction that enhances perturbation. Thus, if we remove some energy from the outer region and give it to the central region, the central region becomes cooler than the outer region, and therefore heat starts to flow inwards. This inward heat flow results in more cooling.

The reason why the central region cools when it gets more thermal energy is that it expands. As a result of the expansion, the gravitational attraction becomes weaker, and the central region can expand more. Through this additional expansion the central region lowers its temperature. In this case, however, the final state is another isothermal state with the same  $E_t$  but a smaller value of  $D$ . Since the minimum value of  $E_t$  corresponds to  $D = 709$ , for any other value of  $D$ , there is one value of  $D$  with same  $E_t$  which is thermally stable.

If the initial perturbation removes the thermal energy from the central region and deposits it to the outer region, the central region becomes hotter, and the heat starts to flow outwards. In this case, there is no stable final state, and the central region continues to contract and becomes hotter indefinitely. Exactly what happens depends on the efficiency of the heat transport.

Roughly speaking, if the thermal timescale is shorter in the center than in the outer region, as the central region contracts, the timescale becomes even shorter. In this case, the central density can reach infinity in a finite time. Of course, in this case, the mass of the region with infinite density should become zero. On the other hand, if the timescale is longer in the center, the timescale becomes even longer as the contraction proceeds, and the overall system contracts slowly. In the case of an ideal gas,

the timescale is determined by the mechanism of the heat transfer.

In the case of stellar systems, heat transfer proceeds through close encounters between two stars. Since the rate of close encounters is directly proportional to the number density of stars, the thermal relaxation timescale is shorter for the central region, and becomes shorter as the central density goes up. This phenomenon is called “gravothermal catastrophe”.

This gravothermal catastrophe occurs in real stellar systems without the adiabatic wall, since real stellar systems always have the temperature decreasing outward, which naturally drives the outward heat flux. Of course, the distribution function is not Maxwellian and one cannot strictly define the temperature. Even so, comparisons between numerical simulations with different approaches have shown that this picture gives a very good description.

Hachisu et al. [9] performed a simulation of the evolution of the self-gravitating gas system with several different forms of heat conductivity. They found that if the heat conductivity is such that the central thermal timescale is shorter than that of the outside region, the evolution of the system would become self-similar. In other words, the central region continues to shrink, leaving a power-law halo behind it. The mass of the core decreases in time and the central density reaches infinity within a finite time.

In the following, we give a simple explanation of this power-law behavior, following the description by Lynden-Bell and Eggleton [14]. Formally, a self-similar solution for a physical quantity  $y$  as a function of radius  $r$  and time  $t$  is expressed as

$$y(r, t) = y_0(t) y_*[r/r_0(t)] . \quad (2)$$

We can set  $y_*(0) = 1$  without loss of generality. We can assume that in the limit of  $r \rightarrow \infty$  there is no evolution, since the thermal timescale is longer at the outskirts. We can further assume that functions  $r_0$  and  $y_0$  are powers of the time  $t$ , since otherwise the self-similar solution cannot be constructed. Thus, if we express

$$r_0 = (t_0 - t)^\beta , \quad (3)$$

and

$$y_0 = (t_0 - t)^\gamma , \quad (4)$$

we have

$$y_0 = r_0^{\gamma/\beta} . \quad (5)$$

The ratio between the gravitational binding energy of the core and the thermal energy of the core should be con-

stant. Therefore, we have

$$\sigma^2 \propto \frac{GM_c}{r_c} \sim \rho_0 r_0^2 . \quad (6)$$

If we express  $\rho_0$  as

$$\rho_0 = r_0^\alpha , \quad (7)$$

we have

$$r_0 = (t_0 - t)^{2/(6+\alpha)} . \quad (8)$$

Lynden-Bell and Eggleton [14] numerically obtained the self-similar solution for the gaseous model with heat conductivity which they believed would mimic the radial energy transfer in an *N*-body system, and found that the self-similar solution has the characteristic power law index of

$$\rho = r^{-2.21} . \quad (9)$$

Henon [11] demonstrated that the *N*-body system would exhibit core collapse using Monte-Carlo calculation with 1000 shells. At that time, *N*-body simulation did show some collapse-like behavior, but it was difficult to see whether the collapse is really self-similar or not because of the limitation in the number of particles. The most beautiful demonstration of the self-similar nature of the collapse is by Cohn [6], who is the first to use the direct integration of the Fokker–Planck equation in the study of the thermal evolution of the globular clusters. He found the power index to be  $-2.23$ , which is strikingly close to the value obtained by the gas model calculation of Lynden-Bell and Eggleton [14].

In the study of the gravothermal catastrophe and self-similar solution, *N*-body simulation did not play a major role. This is partly because more approximate methods, such as the gaseous models and Fokker–Planck calculations, gave reasonable results, and partly because computer power available was quite limited in the 1980s.

### Binary Formation and Gravothermal Oscillation

In the continuous limit, the central density diverges in a finite time, and at that time the core mass becomes zero. Since real stellar systems consist of stars with finite mass and finite size, the central density cannot reach infinity. One possibility is that stars start to physically collide, but there is another possibility.

When the central density becomes very high, the cross-section of the three-body close encounter, which causes the formation of a bound binary, becomes non-negligible.

The binding energy of the binary is transferred to the kinetic energy of the third particle and the center-of-mass motion of the binary. This is essentially the same as the nuclear fusion reaction. The difference is that in this case the interaction which makes the two stars bound is the same gravitational force which makes the entire system bound, and that means there is no ground state for a binary star. Thus, in theory, the binding energy of a binary star can become arbitrarily large. In practice, in real star clusters, there is a practical limit for the binding energy of a binary star. When a binary star interacts with another star, its center-of-mass motion acquires some energy, which is typically a fraction of its binding energy. When the kinetic energy of the center-of-mass motion becomes large enough, the binary would be ejected out the cluster. In fact, this ejection of the binary itself and the other star is the main channel which heats the central region of a star cluster.

This energy production by binaries halts the collapse. The collapse of the core is driven by the outward heat flux. If a sufficient amount of energy is generated in the core, the collapse is halted. In the case of a star, gravitational contraction of the star is halted by the energy production from the nuclear fusion reaction, and it becomes a main-sequence star.

In the case of a star cluster, it is possible to construct a steady-state solution quite similar to the main-sequence stage of a star [8,10]. However, it turned out that this steady-state solution is again thermally unstable [26]. This instability drives what is now called “gravothermal oscillation”. This oscillation was first found in a gas sphere model, but then confirmed with Fokker–Planck

(FP) calculations [7], and later with direct  $N$ -body simulations [16].

Figure 2 shows the time variation of the central density for simulations with 2–32k particles. The time is scaled so that the initial thermal relaxation time is the same for all runs. The curves are shifted vertically.

From Fig. 2 we can see that the oscillation is not strictly periodic. One reason is simply that the binary formation and its interaction with other stars is a stochastic process. The other reason is that, even in the continuous limit with a smooth and deterministic heating term, this oscillation is chaotic. Note that here what exhibits the chaotic behavior is the central density of the whole  $N$ -body system, which is a macroscopic variable of a system with large degrees of freedom.

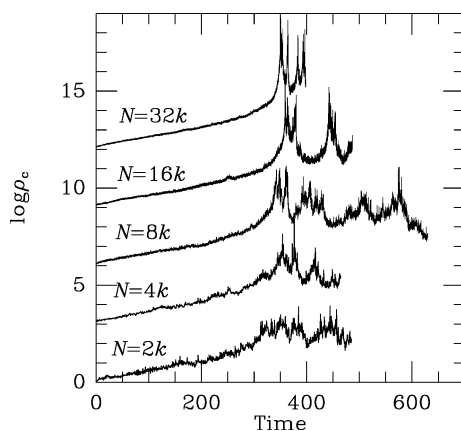
## Numerical Methods

In this section, we discuss the numerical methods used to obtain the solution of gravitational  $N$ -body problem. The problem itself is rather simple. We integrate the equation of motion for stars, using some numerical methods for the initial-value problem of ordinary differential equations. The calculation cost of the mutual gravitation interaction is  $O(N^2)$  and dominate the calculation cost.

In practice, however, what is done is far more complex for a variety of reasons. Here, let us overview the approaches used.

First of all, for some problems, it is not necessary to calculate the interaction of all  $N^2/2$  pairs at each timestep. There are several algorithms which reduce the calculation cost from  $O(N^2)$  to  $O(N \log N)$  or even  $O(N)$ . These algorithms include particle-mesh schemes, Barnes-Hut tree algorithm, and the Fast Multipole Method.

Secondly, it is not practical to apply the same timestep to all stars. While two stars undergo close encounters, their distance can become arbitrary small, depending on the impact parameter and the relative velocity. They have to be integrated with the adaptive timestep which can resolve their relative orbit. Moreover, as we discussed in Sect. “Statistical Description”, many stellar systems tend to develop high-density cores, in which the stars orbit on a timescale much shorter than that of average stars. Also, in many stellar systems, interactions between binary stars (two stars orbiting around each other) and other stars play an important role, and the evolution of the binary orbit due to the interaction with other stars must be accurately integrated. The orbital timescale of binaries can become as small as a fraction of a second, while stars in globular clusters or galaxies have the orbital timescale much longer than one million years. Thus, it is clearly necessary to apply some



Stellar Dynamics,  $N$ -body Methods for, Figure 2

The time variation of the central density of simulated star clusters. (Reproduced from [16])

algorithm with which we can change the timestep of individual stars individually.

Finally, an important characteristic of a stellar system is, from both the theoretical and practical point of views, that it is a Hamiltonian system with various symmetry and conservation laws. Numerical schemes which make use of these characteristics play important roles.

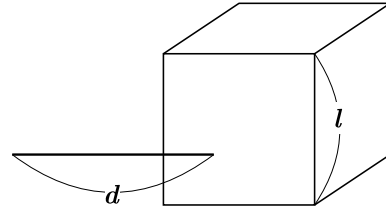
In the following, we briefly overview these schemes.

### Fast Interaction Calculations

The basic idea of the tree code [4] is to replace the force from a group of distant particles with the force from their center of mass or by a multipole expansion. To ensure accuracy, we make groups for distant particles large and groups for nearby particles small.

We use a tree structure to construct the appropriate grouping for each particle. Before calculating the forces on particles, we first organize particles into a tree structure. Barnes and Hut [4] used an oct-tree based on the recursive subdivision of a cube into eight subcubes. We stop the recursive subdivision if the cube has only one particle (or no particles). See [15] for details concerning an efficient tree construction algorithm. Figure 3 shows the Barnes–Hut tree in two dimensional space.

After the tree is constructed, for each node of the tree, which corresponds to a cube of a certain size, we calculate the coefficient of the multipole expansion of the gravita-



Stellar Dynamics, *N*-body Methods for, Figure 4  
Opening criterion for tree traversal

tional force exerted by particles in that cube. The fast algorithm was described in [12].

The force calculation is expressed as a recursive procedure. To calculate the force on a particle, we start from the root node, which corresponds to the total system. We calculate the distance between the node and the particle ( $d$ ) and compare it with the size of the node ( $l$ ). If they satisfy the convergence criterion

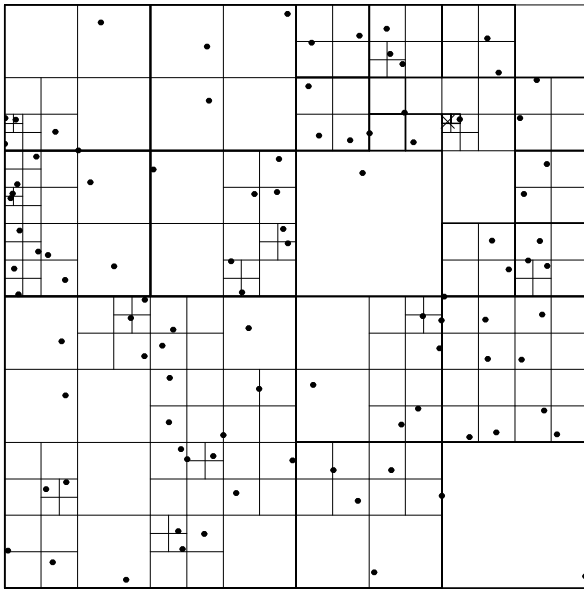
$$\frac{l}{d} < \theta, \quad (10)$$

where  $\theta$  is the accuracy parameter, we calculate the force from that node to the particle using the coefficients of the multipole expansion. If criterion (10) is not satisfied, the force is calculated as a summation of the forces from eight sub-nodes.

Usually, we use the distance between the particle and the center of mass of the node to determine whether the force is accurate enough. When  $\theta$  is very large, this criterion can cause unacceptably large error [21]. For most calculations, however, such a pathological situation is not realized.

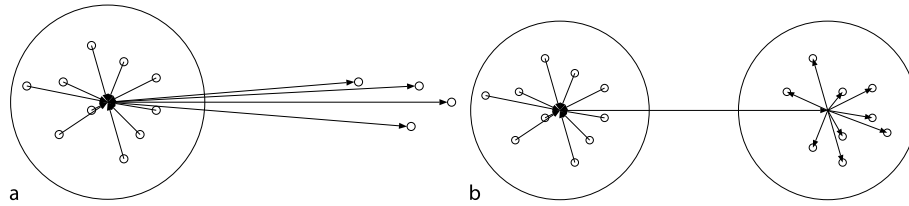
The fast multipole algorithm (FMM) is in some sense a natural extension of the tree algorithm. In the tree algorithm, we use the multipole expansion to express the gravitational force for a group of stars with other stars sufficiently far away (Fig. 5a). In FMM, the forces on another group of stars are expressed in terms of one spherical harmonics expansion, and the forces on these stars are obtained by evaluating the expansion at the locations of stars. The tree-based recursive approach is now applied to both the path to calculate multipole expansions and the path to calculate the spherical harmonics expansions (usually called local expansions). This treatment removes the  $\log N$  term in the calculation cost of the tree algorithm, since the cells of a given size do not directly interact with stars. They interact only with the cells of the same or similar sizes.

After the invention of the tree algorithm and FMM in the late 1980s, a large number of works followed on various topics such as the theoretical analysis of the algo-



Stellar Dynamics, *N*-body Methods for, Figure 3  
Barnes–Hut tree in two dimensions





**Stellar Dynamics, *N*-body Methods for, Figure 5**  
**Tree algorithm (a) and FMM (b)**

rithms, efficient implementation, a fast method to translate multipole and local expansions, parallel implementation, etc. Some of these are covered in [19].

In many other fields of computational science, the Poisson equation associated with the distribution of particles is numerically solved using FFT. If the number of grid points one needs to resolve the potential field is of the same order or less than the number of particles, it is highly advantageous to use FFT, since the calculation cost of FFT is  $O(M \log M)$ , where  $M$  is the number of grid points, with relatively small coefficient. However, FFT requires that the grid is regular with equal spacing. Thus, FFT alone is not very suitable for astronomical simulations, where small-scale structures develop through gravitational instability. Recently, the combination of the tree algorithm and FFT potential solver has become the standard scheme for simulations which requires periodic boundary conditions (see, e. g., Yoshikawa and Fukushige [27]).

### Individual Timestep Algorithms

The individual timestep scheme [1,2] has been the only algorithm that can be used for simulations of gravitational many-body systems, such as open clusters, globular clusters and a system of planetesimals. In a simulation of these systems, we are interested in the change of orbit of each particle due to gravitational encounters with other particles, and the evolution of the total system driven by such changes. In that sense, simulation of these systems is similar to the molecular dynamics simulation, in which we are interested in the thermodynamical process.

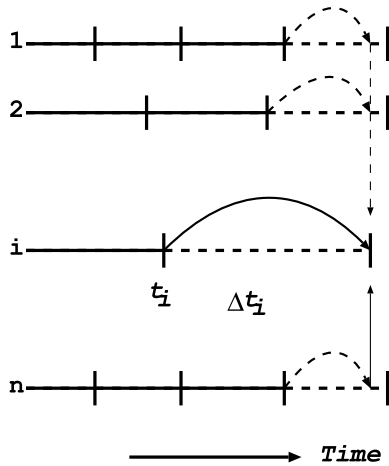
In simulations of these collisional systems, we need to follow the changes of the orbit of a star due to individual encounters with a reasonable accuracy, since the encounters drive the evolution of the system. Therefore, we cannot use the softening parameter to simulate the evolution of these systems. In the study of the galaxies, for example, we can modify the  $1/r$  potential to  $1/\sqrt{r^2 + \epsilon^2}$ , with small constant  $\epsilon$ . This softened potential makes it possible to use a constant and global timestep. When we use a strict  $1/r$  potential, the force between two particles changes very rapidly when two particles undergo a close encounter.

Therefore, during a close encounter, the timestep should be sufficiently small to resolve this rapid change. Roughly speaking, the timestep is determined by the distance to the nearest neighbor. Therefore, if we integrate the system in lockstep, the timestep is determined by the pair of particles with minimum separation. Even in a nearly homogeneous system, the minimum separation is proportional to  $N^{-2/3}$ , and the dependence on  $N$  is stronger when the system is highly inhomogeneous [17].

To reduce the total calculation cost, Aarseth [1] developed an algorithm which assigns each particle its own timestep. In this individual timestep algorithm, each particle adjusts its timestep so that it satisfies the required accuracy. Thus, when two particles undergo a close encounter, although the timesteps of these particles shrink as required, the timesteps of other particles remain long. Makino and Hut [17] showed that the individual timestep scheme is faster than a shared timestep scheme by a factor in the range of  $O(N^{1/3})$  and  $O(N)$ , depending on the distribution of particles.

In the individual timestep algorithm, each particle has its own timestep, and therefore its own time. To calculate the force on a particle due to other particles, we must know their positions at the time of the particle for which we calculate the force. To calculate the position of a particle at a time different from the time of the particle, Aarseth used a third order polynomial extrapolation. This polynomial is evaluated each time a pairwise force is calculated, therefore the calculation cost of the polynomial evaluation is of the same order as that of the force calculation itself. The basic algorithm looks like the following in the case of Aarseth's program, which uses a predictor-corrector scheme:

- (a) Select particle  $i$  with a minimum  $t_i + \Delta t_i$ . Set the global time ( $t$ ) to be this minimum,  $t_i + \Delta t_i$ .
- (b) Predict the positions of all the particles at time  $t$  using the extrapolation polynomial.
- (c) Calculate the acceleration ( $a_i$ ) for particle  $i$  at time  $t$ , using the predicted positions.
- (d) Apply the corrector for the position and velocity of particle  $i$ .
- (e) Go back to step (a).



Stellar Dynamics, *N*-body Methods for, Figure 6  
Schematic description of the individual timestep algorithm

Figure 6 illustrates how the individual timestep scheme works. When particle  $i$  is integrated from  $t_i$  to  $t_i + \Delta t_i$ , the positions of all other particles are predicted at that time, and forces from those particles are calculated using these predicted positions.

Not all time integration schemes can be used with this individual timestep, since the predicted values of the positions of all other particles need to be calculated at the time of the particle to be integrated. Most integration schemes do not provide the solutions at arbitrary points in time. For example, there is no simple way to obtain the approximate solution at the intermediate time with the usual Runge–Kutta schemes or extrapolation schemes.

For the treatment of binaries and small- $N$  systems which can be formed in larger stellar systems, see Aarseth [3].

### Special Numerical Methods

A stellar system is a Hamiltonian system, and therefore the symplectic algorithms [22] can in principle be applied. These schemes show behavior much better than that of traditional schemes, and can dramatically reduce the integration error for long calculations. However, in practice there are a number of difficulties. One is that with symplectic schemes we cannot change timesteps [24]. There has been a number of works to make the symplectic scheme effectively work as variable-timestep scheme. Most of them rely on the partition of the potential energy term of the Hamiltonian into multiple parts with different timescales, and apply different timesteps to different parts of the potential term [20,23]. A very different approach is to use time-symmetric scheme instead of symplectic

schemes. Symmetric schemes offer most of the practical advantages of symplectic schemes, and remain symmetric if we change the timestep. Here, the timestep must be calculated in such a way that does not destroy the time symmetry [13]. Recently, a way to combine this symmetric algorithm and the individual timestep algorithm was proposed [18].

### Future Directions

In Sect. “Statistical Description” we over viewed the evolution of a system of point-mass particles. When we try to understand the evolution of real stellar systems such as globular clusters, open clusters and galactic nuclei, we need to take into account various effects which were neglected in Sect. “Statistical Description”. For example, real stars are not point masses but have finite radii, and can physically collide with each other. In the case of normal stars like our sun in a galaxy, the chance that it will collide with its neighbors before the end of its lifetime is negligible. However, the number density of stars in the central regions of globular clusters or other massive star clusters can reach more than  $10^6/\text{pc}^3$ , or about a million times that of the solar neighborhood. In such circumstances the collision is not rare. Also, individual stars have different masses. Even in the case of thermal equilibrium, stars with different masses have different spacial distribution, since the kinetic energy per unit mass is different. Thus, massive stars tend to lose their kinetic energy and segregate to the central region of the cluster, resulting in strong enhancement of collision rates between most massive stars in star clusters. To make things even more complex, each star evolves from the Main Sequence to giant, and depending on their masses, goes through Type II supernova and finally becomes a white dwarf, a neutron star, or a black hole. These stellar evolutions and stellar collisions, coupled with the stellar dynamics, makes stellar systems very interesting and an important research subject.

The ultimate research direction is to model all of these processes, stellar evolution, stellar collisions, stellar dynamics, and at some point the initial gas dynamics of star cluster formation, in one single unified simulation. From the viewpoint of the computational cost, such a simulation is not impossible, since the calculation cost of stellar dynamics is high and others are relatively minor. However, from the viewpoint of software engineering and physical modeling, we do not really know how we can actually develop and maintain a simulation program which can handle gravitational interaction between stars, evolution of individual stars, physical collisions and tidal interaction between stars, and other more complex things like the evo-

lution of binary stars. MODEST (Modeling Dense Stellar systems)<sup>1</sup> is one of such efforts to develop a simulation code, or a loosely coupled collection of codes, to handle complex systems.

## Bibliography

1. Aarseth SJ (1963) MN 126:223
2. Aarseth SJ (1985) In: Blackbill JU, Cohen BI (eds) Multiple Time Scales. Academic Press, New York, pp 377–418
3. Aarseth SJ (2003) Gravitational N-Body Simulations. Cambridge University Press, Cambridge, pp 430
4. Barnes J, Hut P (1986) Nature 324:446
5. Binney J, Tremaine S (2008) Galactic Dynamics, 2nd edn. Princeton University Press, Princeton
6. Cohn H (1980) ApJ 242:765
7. Cohn H, Hut P, Wise M (1989) ApJ 342:814
8. Goodman J (1984) ApJ 280:298
9. Hachisu I, Nakada Y, Nomoto K, Sugimoto D (1978) Prog Theor Phys 60:393
10. Heggie DC (1984) MN 206:179
11. Henon M (1971) Astrophys Space Sci 13:284
12. Hernquist L (1990) J Comp Phys 87:137
13. Hut P, Makino J, McMillan S (1995) ApJL 443:L93
14. Lynden-Bell D, Eggleton PP (1980) MN 191:483
15. Makino J (1990) J Comput Phys 87:148
16. Makino J (1996) Dynamical Evolution of Star Clusters. In: Hut P, Makino J (eds) Proc IAU Symp 174. Kluwer, Amsterdam, pp 151–160
17. Makino J, Hut P (1988) ApJS 68:833
18. Makino J, Hut P, Kaplan M, Saygin H (2006) New Astron 12:124
19. Pfalzner S, Gibbon P (1996) Many-Body Tree Methods in Physics. Cambridge University Press, Cambridge
20. Preto M, Tremaine S (1999) AJ 118:2532
21. Salmon JK, Warren MS (1994) J Comp Phys 111:136
22. Sanz-Serna JM, Calvo MP (1994) Numerical Hamiltonian Problems. Chapman and Hall, London
23. Skeel RD, Biesiadecki JJ (1994) Ann Numer Math 1:191
24. Skeel RD, Gear CW (1992) Physica D 60:311
25. Spitzer Jr L (1987) Dynamical Evolution of Globular Clusters. Princeton University Press, Princeton
26. Sugimoto D, Bettwieser E (1983) MN 204:19P
27. Yoshikawa K, Fukushima T (2005) PASJ 57:849

## Stochastic Games

EILON SOLAN

The School of Mathematical Sciences,  
Tel Aviv University, Tel Aviv, Israel

## Article Outline

Glossary

Definition of the Subject

Strategies, Evaluations and Equilibria

Zero-Sum Games

Multi-Player Games

Correlated Equilibrium

Imperfect Monitoring

Algorithms

Additional and Future Directions

Acknowledgments

Bibliography

## Glossary

**A stochastic game** A repeated interaction between several participants in which the underlying state of the environment changes stochastically, and it depends on the decisions of the participants.

**A strategy** A rule that dictates how a participant in an interaction makes his decisions as a function of the observed behavior of the other participants and of the evolution of the environment.

**Evaluation of stage payoffs** The way that a participant in a repeated interaction evaluates the stream of stage payoffs that he receives (or stage costs that he pays) along the interaction.

**An equilibrium** A collection of strategies, one for each player, such that each player maximizes (or minimizes, in case of stage costs) his evaluation of stage payoffs given the strategies of the other players.

**A correlated equilibrium** An equilibrium in an extended game in which at the outset of the game each player receives a private signal, and the vector of private signals is chosen according to a known joint probability distribution. In the extended game, a strategy of a player depends, in addition to past play, on the signal he received.

## Definition of the Subject

Stochastic games, first introduced by Shapley [60], model dynamic interactions in which the environment changes in response to the behavior of the players. Formally, a stochastic game is a tuple  $G = \langle N, S, (\mathcal{A}_i, A_i, u_i)_{i \in N}, q \rangle$  where

- $N$  is a set of players.
- $S$  is a state space. If  $S$  is uncountable, it is supplemented with a  $\sigma$ -algebra of measurable sets.
- For every player  $i \in N$ ,  $\mathcal{A}_i$  is a set of actions for that player, and  $A_i: S \rightarrow \mathcal{A}_i$  is a set-valued (measurable) function that assigns to each state  $s \in S$  the set of actions  $A_i(s)$  that are available to player  $i$  in state  $s$ . If  $\mathcal{A}_i$  is uncountable, it is supplemented with a  $\sigma$ -algebra of measurable sets. Denote  $SA = \{(s, a): s \in S, a =$

<sup>1</sup><http://www.manybody.org/modest/>

$(a_i)_{i \in N}, a_i \in A_i(s) \ \forall i \in N\}$ . This is the set of all possible *action profiles*.

- For every player  $i \in N$ ,  $u_i: SA \rightarrow \mathbf{R}$  is a (measurable) *stage payoff function* for player  $i$ .
- $q: SA \rightarrow \Delta(S)$  is a (measurable) *transition function*, where  $\Delta(S)$  is the space of probability distributions over  $S$ .

The game starts at an initial state  $s^1$ , and is played as follows. At each stage  $t \in \mathbf{N}$ , each player  $i \in N$  chooses an action  $a_i^t \in A_i(s^t)$ , receives the stage payoff  $u_i(s^t, a^t)$ , where  $a^t = (a_i^t)_{i \in N}$ , and the game moves to a new state  $s^{t+1}$  that is chosen according to the probability distribution  $q(\cdot | s^t, a^t)$ .

A few comments are in order.

1. A stochastic game lasts infinitely many stages. However, the model also captures finite interactions (of length  $t$ ), by assuming the play moves, at stage  $t$ , to an absorbing state with payoff 0 to all players.
2. In particular, by setting  $t = 1$ , we see that stochastic games are a generalization of matrix games (games in normal form/games in normal form), which are played only once.
3. Stochastic games are also a generalization of repeated games, in which the players play the same matrix game over and over again. Indeed, a repeated game is equivalent to a stochastic game with a single state.
4. Stopping games are also a special case of stochastic games. In these games every player has two actions in all states, *continue* and *quit*. as long as all players choose *continue* the stage payoff is 0; once at least one player chooses *quit* the game moves to an absorbing state.
5. Markov decision problems (see, e.g., [49]) are stochastic games with a single player.
6. The transition function  $q$  governs the evolution of the game. It depends on the actions of all players and on the current state, so that all the players influence the evolution of the game.
7. The payoff function  $u_i$  of player  $i$  depends on the current state as well as on the actions chosen by all players. Thus, a player's payoff depends not only on that player's choice, but also on the behavior of the other players.
8. Though we refer to the functions  $(u_i)_{i \in N}$  as "stage payoffs", with the implicit assumption that each player tries to maximize his payoff, in some applications these functions describe a stage *cost*, and then the implicit assumption is that each player tries to minimize his cost.
9. The action of a player at a given stage affects both his stage payoff and the evolution of the state variable, thereby affecting his future payoffs. These two, sometimes contradicting effects make the optimization problem of the players quite intricate, and the analysis of the game challenging.
10. The players receive a stage payoff at each stage. So far we did not mention how the players evaluate the infinite stream of stage payoffs that they receive, nor did we say what is their information at each stage: Do they observe the current state? Do they observe the actions of the other players? These issues will be discussed later.
11. The actions that are available to the players at each stage, the payoff functions, and the transition function, all depend on the current state, and not on past play (that is, past states that the game visited, and past actions that the players chose). This assumption is without loss of generality. Indeed, suppose that the actions available to the players at each stage, the payoff functions, and the transition function, all depend on past play, as well as on the current state. For every  $t \in \mathbf{N}$  let  $H_t$  be the set of all possible *histories* of length  $t$ , that is, all sequences of the form  $(s^1, a^1, s^2, a^2, \dots, s^t)$ , where  $s^k \in S$  for every  $k = 2, 3, \dots, t$ ,  $a^k = (a_i^k)_{i \in N}$  and  $a_i^k$  is an available action to player  $i$  at stage  $k$ , for every  $k = 1, 2, \dots, t-1$ . Then the game is equivalent to a game with state space  $H := \bigcup_{t \in \mathbf{N}} H_t$ , in which the state variable captures past play, and the state at stage  $t$  lies in  $H_t$ . In the new game, the sets of available actions, the payoff function, and the transition function, depend on the current state rather than on all past play.

The interested reader is referred to [20,42,72] for further reading on stochastic games. We now provide a few applications.

*Example 1 (Capital Accumulation ([7,18,19,34,45]))* Two (or more) agents jointly own a natural resource or a productive asset; at every period they have to decide the amount of the resource to consume. The amount that is not consumed grows by a known (or an unknown) fraction. Such a situation occurs, e.g., in fishery: Fishermen from various countries fish in the same area, and each country sets a quota for its fishermen. Here the state variable is the current amount of resource, the action set is the amount of resource to be exploited in the current period, and the transition is influenced by the decisions of all the players, as well as possibly by the random growth of the resource.

**Example 2 (Taxation ([14,48]))** A government sets a tax rate at every period. Each citizen decides at every period how much to work, and, from the total amount of money he or she has, how much to consume; the rest is saved for the next period, and grows by a known interest rate. Here the state is the amount of savings each citizen has, the stage payoff of a citizen depends on the amount of money that he consumed, on the amount of free time he has, and on the total amount of tax that the government collected. The stage payoff of the government may be the average stage payoff of the citizens, the amount of tax collected, or a mixture of the two.

**Example 3 (Communication Network [58])** A single-cell system with one receiver and multiple uplink transmitters share a single, slotted, synchronous classical collision channel. Assume that all transmitted packets have the same length, and require one time unit, which is equal to one time slot, for transmission. Whenever a collision occurs, the users attempt to retransmit their packets in subsequent slots to resolve collision for reliable communication.

Here a state lists all relevant data for a given stage: e. g., the number of packets waiting at each transmitter, or the length of time each has been waiting to be transmitted. The players are the transmitters, and the action of each transmitter is which packet to transmit, if any. The stage cost may depend on the number of time slots that the transmitted packet waited, on the number of packets that have not been transmitted at that period, and possibly on additional variables. The transition depends on the actions chosen by the players, but it has a stochastic component, which captures the number of new packets that arrive at the various transmitters during every time slot.

**Example 4 (Queues [1])** Individuals that require service have to choose whether to be served by a private slow service provider, or by a powerful public service provider. This situation arises, e. g., when jobs can be executed on either a slow personal computer or a fast mainframe. Here a state lists the current load of the public and private service providers, and the cost is the time to be served.

The importance of stochastic games stems from the wide range of applications they encompass. Many repeated interactions can be recast as stochastic games; the wide range of theoretical results that have been obtained provide insights that can help in analyzing specific situations and suggesting proper behavior to the participants. In certain classes of games algorithms that have been developed may be used to calculate such behavior.

## Strategies, Evaluations and Equilibria

So far we have not described the information that the players have at each stage. In most of the chapter we assume that the players have complete information of past play; that is, at each stage  $t$ , they know the sequence  $s^1, a^1, s^2, a^2, \dots, s^t$  of states that were visited in the past (including the current state) and the actions that were chosen by all players. This assumption is too strong for most applications, and in the sequel we will mention the consequences of its relaxation.

Since the players observe past play, a *pure strategy* for player  $i$  is a (measurable) function  $\sigma_i$  that assigns to every finite history  $(s^1, a^1, s^2, a^2, \dots, s^t)$  an action  $\sigma_i(s^1, a^1, s^2, a^2, \dots, s^t) \in A_i(s^t)$ , with the interpretation that, at stage  $t$ , if the finite history  $(s^1, a^1, s^2, a^2, \dots, s^t)$  occurred, player  $i$  plays the action  $\sigma_i(s^1, a^1, s^2, a^2, \dots, s^t)$ . If the player does not know the complete history, then a strategy for player  $i$  is a function that assigns to every possible information set, an action that is available to the player when the player has this information. A *mixed strategy* for player  $i$  is a probability distribution over the set of his pure strategies. The space of mixed strategies of player  $i$  is denoted by  $\sigma_i$ .

A simple class of strategies is the class of *stationary* strategies; a strategy  $\sigma_i$  for player  $i$  is *stationary* if  $\sigma_i(s^1, a^1, s^2, a^2, \dots, s^t)$  depends only on the current state  $s^t$ , and not on past play  $s^1, a^1, s^2, a^2, \dots, a^{t-1}$ . A stationary strategy of player  $i$  can be identified with an element  $x = (x_s)_{s \in S} \in \times_{s \in S} \Delta(A_i(s))$ , with the interpretation that player  $i$  plays the mixed action  $x_s$  whenever the current state is  $s$ . Denote by  $X_i = \times_{s \in S} \Delta(A_i(s))$  the space of stationary strategies of player  $i$ .

There are three common ways to evaluate the infinite stream of payoffs that the players receive in a stochastic game: The *finite-horizon evaluation*, in which a player considers the average payoff during the first  $T$  stages, the *discounted evaluation*, in which a player considers the discounted sum of his stage payoffs, and the *limsup evaluation*, in which a player considers the limsup of his long-run average payoffs. We now formally define these evaluations.

Every profile  $\sigma = (\sigma_i)_{i \in N}$  of mixed strategies, together with the initial state, induces a probability distribution  $\mathbf{P}_{s_1, \sigma}$  over the space of infinite plays  $H_\infty := \text{SAN}$ . We denote the corresponding expectation operator by  $\mathbf{E}_{s_1, \sigma}$ .

**Definition 5** Let  $\sigma$  be a profile of mixed strategies. For every finite horizon  $T \in \mathbf{N}$ , the  *$T$ -stage payoff* under  $\sigma$  for player  $i$  is

$$\gamma_i^T(s_1, \sigma) := \mathbf{E}_{s_1, \sigma} \left[ \frac{1}{T} \sum_{t=1}^T u_i(s^t, a^t) \right].$$



For every discount factor  $\lambda \in (0, 1]$ , the  $\lambda$ -discounted payoff under  $\sigma$  for player  $i$  is

$$\gamma_i^\lambda(s_1, \sigma) := \mathbf{E}_{s_1, \sigma} \left[ \lambda \sum_{t=1}^{\infty} (1 - \lambda)^{t-1} u_i(s^t, a^t) \right].$$

The *limsup payoff* under  $\sigma$  for player  $i$  is

$$\gamma_i^\infty(s_1, \sigma) := \mathbf{E}_{s_1, \sigma} \left[ \limsup_{T \rightarrow \infty} \frac{1}{T} \sum_{t=1}^T u_i(s^t, a^t) \right].$$

The  $T$ -stage payoff captures the situation in which the interaction lasts exactly  $T$  stages. The  $\lambda$ -discounted evaluation captures the situation in which the game lasts “many” stages, and the player discounts stage payoffs – it is better to receive \$1 today than tomorrow. The limsup payoff also captures the situation in which the game lasts “many” stages, but here the player does not discount his payoffs, and the payoff at each given stage is insignificant as compared to the payoff in all other stages. Equivalently, one could consider the liminf payoff in which the player considers the liminf of the long-run average payoffs.

As usual, an equilibrium is a vector of strategies such that no player can profit by a unilateral deviation. For every player  $i$  and every strategy profile  $\sigma = (\sigma_i)_{i \in N}$  we denote the strategy profile of all other players, except player  $i$ , by  $\sigma_{-i} = (\sigma_j)_{j \neq i}$ .

**Definition 6** Let  $\varepsilon \geq 0$ . A profile of strategies  $\sigma$  is a  $T$ -stage  $\varepsilon$ -equilibrium if

$$\gamma_i^T(s_1, \sigma) \geq \gamma_i^T(s_1, \sigma'_i, \sigma_{-i}) - \varepsilon, \\ \forall s_1 \in S, \forall i \in N, \forall \sigma'_i \in \Sigma_i.$$

It is a  $\lambda$ -discounted  $\varepsilon$ -equilibrium if

$$\gamma_i^\lambda(s_1, \sigma) \geq \gamma_i^\lambda(s_1, \sigma'_i, \sigma_{-i}) - \varepsilon, \\ \forall s_1 \in S, \forall i \in N, \forall \sigma'_i \in \Sigma_i.$$

It is a *limsup  $\varepsilon$ -equilibrium* if

$$\gamma_i^\infty(s_1, \sigma) \geq \gamma_i^\infty(s_1, \sigma'_i, \sigma_{-i}) - \varepsilon, \\ \forall s_1 \in S, \forall i \in N, \forall \sigma'_i \in \Sigma_i.$$

The payoff that corresponds to an  $\varepsilon$ -equilibrium, that is, either one of the quantities  $\gamma_i^T(s_1, \sigma)$ ,  $\gamma_i^\lambda(s_1, \sigma)$  and  $\gamma_i^\infty(s_1, \sigma)$ , is called an  $\varepsilon$ -equilibrium payoff at the initial state  $s_1$ .

As we will see below, when both state and action spaces are finite, a  $T$ -stage and a  $\lambda$ -discounted 0-equilibrium exist.

However, when the state or action spaces are infinite such a 0-equilibrium may fail to exist, yet  $\varepsilon$ -equilibria may exist for every  $\varepsilon > 0$ .

As the length of the game  $T$  varies, or as the discount factor  $\lambda$  varies, the equilibrium strategy profile varies as well. A strategy profile that is an  $\varepsilon$ -equilibrium for every  $T$  sufficiently large and for every  $\lambda$  sufficiently small is called a *uniform  $\varepsilon$ -equilibrium*.

**Definition 7** Let  $\varepsilon > 0$ . A strategy profile  $\sigma$  is a *uniform  $\varepsilon$ -equilibrium* if there are  $T_0 \in \mathbf{N}$  and  $\lambda_0 \in (0, 1)$  such that for every  $T \geq T_0$  the strategy profile  $\sigma$  is a  $T$ -stage  $\varepsilon$ -equilibrium, and for every  $\lambda \in (0, \lambda_0)$  it is a  $\lambda$ -discounted  $\varepsilon$ -equilibrium.

If for every  $\varepsilon > 0$  the game has a ( $T$ -stage,  $\lambda$ -discounted, limsup or uniform)  $\varepsilon$ -equilibrium with corresponding payoff  $g_\varepsilon$ , then any accumulation point of  $(g_\varepsilon)_{\varepsilon > 0}$  as  $\varepsilon$  goes to 0 is a ( $T$ -stage,  $\lambda$ -discounted, limsup or uniform) equilibrium payoff.

## Zero-Sum Games

A two-player stochastic game is *zero-sum* if  $u_1(s, a) + u_2(s, a) = 0$  for every  $(s, a) \in SA$ . As in matrix games, every two-player zero-sum stochastic game admits at most one equilibrium payoff at every initial state  $s_1$ , which is termed the *value* of the game at  $s_1$ . Each player's strategy which is part of an  $\varepsilon$ -equilibrium is termed  *$\varepsilon$ -optimal*. The definition of  $\varepsilon$ -equilibrium implies that an  $\varepsilon$ -optimal strategy guarantees the value up to  $\varepsilon$ ; for example, in the  $T$ -stage evaluation, if  $\sigma_1$  is an  $\varepsilon$ -optimal strategy of player 1, then for every strategy of player 2 we have

$$\gamma_1^T(s_1, \sigma_1, \sigma_2) \geq v^T(s_1) - \varepsilon,$$

where  $v^T(s_1)$  is the  $T$ -stage value at  $s_1$ .

In his seminal work, Shapley [60] presented the model of two-player zero-sum stochastic games with finite state and actions spaces, and proved the following.

**Theorem 8 [60]** *For every two-player zero-sum stochastic game, the  $\lambda$ -discounted value at every initial state exists. Moreover, both players have  $\lambda$ -discounted 0-optimal stationary strategies.*

*Proof* Let  $\mathcal{V}$  be the space of all functions  $v: S \rightarrow \mathbf{R}$ . For every  $v \in \mathcal{V}$  define a zero-sum matrix game  $G_s^\lambda(v)$  as follows:

- The action spaces of the two players are  $A_1(s)$  and  $A_2(s)$  respectively.
- The payoff function (that player 2 pays player 1) is

$$\lambda u_1(s, a) + (1 - \lambda) \sum_{s' \in S} q(s' | s, a) v(s').$$

The game  $G_s^\lambda(v)$  captures the situation in which, after the first stage, the game terminates with a terminal payoff  $v(s')$ , where  $s'$  is the state reached after stage 1. Define an operator  $\varphi: \mathcal{V} \rightarrow \mathcal{V}$  as follows:

$$\varphi_s(v) = \text{val}(G_s^\lambda(v)),$$

where  $\text{val}(G_s^\lambda(v))$  is the value of the matrix game  $G_s^\lambda(v)$ . Since the value operator is non-expansive, it follows that the operator  $\varphi$  is contracting:  $\|\varphi(v) - \varphi(w)\|_\infty \leq (1 - \lambda)\|v - w\|_\infty$ , so that this operator has a unique fixed point  $\hat{v}^\lambda$ . One can show that the fixed point is the value of the stochastic game, and every strategy  $\sigma_i$  of player  $i$  in which he plays, after each finite history  $(s^1, a^1, s^2, a^2, \dots, s^t)$ , an optimal mixed action in the matrix game  $G_{s^t}^\lambda(\hat{v}^\lambda)$ , is a  $\lambda$ -discounted 0-optimal strategy in the stochastic game.  $\square$

**Example 9** Consider the following two-player zero-sum game with three states,  $s_0, s_1$  and  $s_2$ ; each entry of the matrix indicates the payoff that player 2 (the column player) pays player 1 (the row player, the payoff is in the middle), and the transitions (which are deterministic, and are denoted at the top-right corner).

|   |             |         |  |             |  |             |
|---|-------------|---------|--|-------------|--|-------------|
|   | L           | R       |  | L           |  | L           |
| T | 0 $s^2$     | 1 $s^1$ |  | 1 $s^1$     |  | 0 $s^0$     |
| B | 1 $s^1$     | 0 $s^0$ |  |             |  |             |
|   | State $s_2$ |         |  | State $s_1$ |  | State $s_0$ |

The states  $s_0$  and  $s_1$  are absorbing: Once the play reaches one of these states it never leaves it. State  $s_2$  is non-absorbing. Stochastic games with a single non-absorbing state are called *absorbing games*. For every  $v = (v_0, v_1, v_2) \in \mathcal{V} = \mathbf{R}^3$  the game  $G_{s_2}^\lambda(v)$  is the following matrix game:

|   |                              |                              |
|---|------------------------------|------------------------------|
|   | L                            | R                            |
| T | $(1 - \lambda)v_2$           | $\lambda + (1 - \lambda)v_1$ |
| B | $\lambda + (1 - \lambda)v_1$ | $(1 - \lambda)v_0$           |

The game  $G_{s_2}^\lambda$

|   |                              |  |                    |
|---|------------------------------|--|--------------------|
|   | L                            |  | L                  |
| T | $\lambda + (1 - \lambda)v_1$ |  | $(1 - \lambda)v_0$ |

The game  $G_{s_1}^\lambda$       The game  $G_{s_0}^\lambda$

The unique fixed point of the operator  $\text{val}(G^\lambda)$  must satisfy

- $\hat{v}_0 = \text{val}(G_{s_0}^\lambda(\hat{v}))$ , so that  $\hat{v}_{s_0}^\lambda = \hat{v}_0 = 0$ ;
- $\hat{v}_1 = \text{val}(G_{s_1}^\lambda(\hat{v}))$ , so that  $\hat{v}_{s_1}^\lambda = \hat{v}_1 = 1$ ;

- $\hat{v}_2 = \text{val}(G_{s_2}^\lambda(\hat{v}))$ . By Theorem 8 both players have a stationary  $\lambda$ -discounted 0-optimal strategy. Denote by  $x$  (resp.  $y$ ) a mixed action for player 1 (resp. player 2) that is part of a  $\lambda$ -discounted 0-optimal strategy at the state  $s_2$ . Since we know that in the fixed point  $\hat{v}_0 = 0$  and  $\hat{v}_1 = 1$ ,  $\hat{v}_2$  must be the unique solution of

$$v_2 = y(1 - \lambda)v_2 + (1 - y) = y,$$

so that  $\hat{v}_{s_2}^\lambda = \hat{v}_2 = (1 - \sqrt{\lambda})/(1 - \lambda)$ . The 0-optimal strategy of player 2 at state  $s_2$  is  $y = \hat{v}_2 = (1 - \sqrt{\lambda})/(1 - \lambda)$ , and the 0-optimal strategy of player 1,  $x = \hat{v}_2 = (1 - \sqrt{\lambda})/(1 - \lambda)$ , can be found by finding his 0-optimal strategy in  $G_{s_2}^\lambda(\hat{v})$ .

Bewley and Kohlberg [11] proved that when the state and action spaces are finite, the function  $\lambda \mapsto v_s^\lambda$ , that assigns to every state  $s$  and every discount factor  $\lambda$  the  $\lambda$ -discounted value at the initial state  $s$ , is a *Puiseux function*, that is, it has a representation  $v_s^\lambda = \sum_{k=K}^\infty a_k \lambda^{k/M}$  that is valid for every  $\lambda \in (0, \lambda_0)$  for some  $\lambda_0 > 0$ , where  $M$  is a natural number,  $K$  is a non-negative integer, and  $(a_k)_{k=K}^\infty$  are real numbers. In particular, the function  $\lambda \mapsto v_s^\lambda$  is monotone in a neighborhood of 0, and its limit as  $\lambda$  goes to 0 exists. This result turned out to be crucial in subsequent study on games with finitely many states and actions.

Shapley's work has been extended to general state and action spaces; for a recent survey see [46]. The tools developed in [46], together with a dynamic programming argument, prove that under proper conditions on the payoff function and on the transitions the two-player zero-sum stochastic game has a  $T$ -stage value.

Maitra and Sudderth [35] proved that the limsup value exists in a very general setup. Their proof follows closely that of Martin [36] for the determinacy of Blackwell games.

The study of the uniform value emanated from an example, called the "Big Match", due to Gillette [28], that was solved by Blackwell and Ferguson [13].

**Example 10** Consider the following stochastic game with two absorbing states and one non-absorbing state.

|   |             |         |  |             |  |             |
|---|-------------|---------|--|-------------|--|-------------|
|   | L           | R       |  | L           |  | L           |
| T | 0 $s^2$     | 1 $s^2$ |  | 1 $s^1$     |  | 0 $s^0$     |
| B | 1 $s^1$     | 0 $s^0$ |  |             |  |             |
|   | State $s_2$ |         |  | State $s_1$ |  | State $s_0$ |

Suppose the initial state is  $s_2$ . As long as player 1 plays  $T$  the play remains at  $s_2$ ; once he plays  $B$  the play moves to either  $s_0$  or  $s_1$ , and is effectively terminated. By finding the fixed point of the operator  $\varphi$  one can show

that the discounted value at the initial state  $s_2$  is  $\frac{1}{2}$ , and a  $\lambda$ -discounted stationary 0-optimal strategy for player 2 is<sup>1</sup>  $[\frac{1}{2}(L), \frac{1}{2}(R)]$ . Indeed, if player 1 plays  $T$  then the expected stage payoff is  $\frac{1}{2}$  and play remains at  $s_2$ , while if player 1 plays  $B$  then the game moves to an absorbing state, and the expected stage payoff from that stage onwards is  $\frac{1}{2}$ . In particular, this strategy guarantees  $\frac{1}{2}$  for player 2 both in the limsup evaluation and uniformly. A  $\lambda$ -discounted 0-optimal strategy for player 1 is  $[\frac{1}{1+\lambda}(T), \frac{\lambda}{1+\lambda}(B)]$ .

What can player 1 guarantee in the limsup evaluation and uniformly? If player 1 plays the stationary strategy  $[x(T), (1-x)(B)]$  that plays at each stage the action  $T$  with probability  $x$  and the action  $B$  with probability  $1-x$ , then player 2 has a reply that ensures that the limsup payoff is 0: If  $x = 1$  and player 2 always plays  $L$ , the payoff is 0 at each stage; if  $x < 1$  and player 2 always plays  $R$ , the payoff is 1 until the play moves to  $s_0$ , and then it is 0 forever. Since player 1 plays the action  $B$  with probability  $1-x > 0$  at each stage, the distribution of the stage in which play moves to  $s_0$  is geometric. Therefore, the limsup payoff is 0, and if  $\lambda$  is sufficiently small, the discounted payoff is close to 0.

One can verify that if player 1 uses a bounded-recall strategy, that is, a strategy that uses only the last  $k$  actions that were played, player 2 has a reply that guarantees that the limsup payoff is 0, and the discounted payoff is close to 0, provided  $\lambda$  is close to 0. Thus, in the limsup payoff and uniformly finite memory cannot guarantee more than 0 in this game (see also [27]).

Intuitively, player 1 would like to condition the probability of playing  $T$  on the past behavior of player 2: If in the past player 2 played the action  $L$  more often than the action  $R$ , he would have liked to play  $T$  with higher probability; if in the past player 2 played the action  $R$  more often than the action  $L$ , he would have liked to play  $B$  with higher probability. Blackwell and Ferguson [13] constructed a family of good strategies  $\{\sigma_1^M, M \in \mathbf{N}\}$  for player 1. The parameter  $M$  determines the amount that the strategy guarantees: The strategy  $\sigma_1^M$  guarantees a limsup payoff and a discounted payoff of  $\frac{M}{2M+1}$ , provided the discount factor is sufficiently low. In other words, player 1 cannot guarantee  $\frac{1}{2}$ , but he may guarantee an amount as close to  $\frac{1}{2}$  as he wishes by choosing  $M$  to be sufficiently large. The strategy  $\sigma_1^M$  is defined as follows: At stage  $t$ , play  $B$  with probability  $\frac{1}{(M+l_t-r_t)^2}$ , where  $l_t$  is the number of stages up to stage  $t$  in which player 2 played  $L$ , and  $r_t$  is the number of stages up to stage  $t$  in which player 2 played  $R$ .

<sup>1</sup>That is, at each stage player 2 plays  $L$  with probability  $\frac{1}{2}$  and  $R$  with probability  $\frac{1}{2}$ .

Since  $r_t + l_t = t - 1$  one has  $r_t - l_t = 2r_t - (t - 1)$ . The quantity  $r_t$  is the total payoff that player 1 received in the first  $t - 1$  stages if player 1 played  $T$  in those stages (and the game was not absorbed). Thus, this total payoff is a linear function of the difference  $r_t - l_t$ . When presented this way, the strategy  $\sigma_1^M$  depends on that total payoff. Observe that as  $r_t$  increases,  $r_t - l_t$  increases as well, and the probability to play  $B$  decreases.

Mertens and Neyman [38] generalized the idea presented at the end of Example 10 to stochastic games with finite state and action spaces.<sup>2</sup>

**Theorem 11** *If the state and action spaces of a two-player zero-sum stochastic game are finite, the game has a uniform value  $v_s^0$  at every initial state  $s \in S$ . Moreover,  $v_s^0 = \lim_{\lambda \rightarrow 0} v_s^\lambda = \lim_{T \rightarrow \infty} v_s^T$ .*

In their proof, Mertens and Neyman describe a uniform  $\varepsilon$ -optimal strategy. In this strategy the player keeps a parameter,  $\lambda_t$ , which is a fictitious discount factor to use at stage  $t$ . This parameter changes at each stage as a function of the stage payoff; if the stage payoff at stage  $t$  is high then  $\lambda_{t+1} < \lambda_t$ , whereas if the stage payoff at stage  $t$  is low then  $\lambda_{t+1} > \lambda_t$ . The intuition is as follows. As mentioned before, in stochastic games there are two forces that influence the player's behavior: He tries to get high stage payoffs, while keeping future prospects high (by playing in such a way that the next stage that is reached is favorable). When considering the  $\lambda$ -discounted payoff there is a clear comparison between the importance of the two forces: The weight of the stage payoff is  $\lambda$  and the weight of future prospects is  $1 - \lambda$ ; the lower the discount factor, the more weight is given to the future. When considering the uniform value (or the uniform equilibrium) the weight of the stage payoff is 0. However, if the player never attempts to receive a high stage payoff, the overall payoff in the game will not be high. Therefore, the player has a fictitious discount factor; if past payoffs are low and they do not meet the expectation, player 1 increases the weight of the stage payoff by increasing the fictitious discount factor; if past payoffs are high, player 1 increases the weight of the future by lowering this fictitious discount factor.

## Multi-Player Games

Takahashi [75] and Fink [21] extended Shapley's [60] result to discounted equilibria in non-zero-sum games.

<sup>2</sup>Mertens and Neyman's [38] result actually holds in every stochastic game that satisfies a proper condition, which is always satisfied when the state and action spaces are finite.

**Theorem 12** *Every stochastic game with finite state and action spaces has a  $\lambda$ -discounted equilibrium in stationary strategies.*

*Proof* The proof utilizes Kakutani's fixed point theorem [31]. Let  $M = \max_{i,s,a} |u_i(s, a)|$  be a bound on the absolute values of the payoffs. Set  $X = \times_{i \in N, s \in S} (\Delta(A_i(s)) \times [-M, M])$ . A point  $x = (x_{i,s}^A, x_{i,s}^V)_{i \in N, s \in S} \in X$  is a collection of one mixed action and one payoff to each player at every state. For every  $v = (v_i)_{i \in N} \in [-M, M]^{N \times S}$  and every  $s \in S$  define a matrix game  $G_s^\lambda(v)$  as follows:

- The action spaces of each player  $i$  is  $A_i(s)$ ;
- The payoff to player  $i$  is

$$\lambda u_i(s, a) + (1 - \lambda) \sum_{s' \in S} q(s' | s, a) v_i(s').$$

We define a set-valued function  $\varphi: X \rightarrow X$  as follows.

- For every  $i \in N$  and every  $s \in S$ ,  $\varphi_{i,s}^A$  is the set of all best responses of player  $i$  to the strategy vector  $x_{-i,s} := (x_{j,s})_{j \neq i}$  in the game  $G_s^\lambda(v)$ . That is,

$$\varphi_{i,s}^A(x, v) := \left\{ \operatorname{argmax}_{y_{i,s} \in \Delta(A_i(s))} \lambda r_i(s, y_{i,s}, x_{-i,s}) + (1 - \lambda) \sum_{s' \in S} q(s' | s, y_{i,s}, x_{-i,s}) v_{i,s'} \right\}.$$

- For every  $i \in N$  and every  $s \in S$ ,  $\varphi_{i,s}^V(x, v)$  is the maximal payoff for player  $i$  in the game  $G_s^\lambda(v)$ , when the other players play  $x_{-i}$ :

$$\varphi_{i,s}^V(x, v) := \max_{y_{i,s} \in \Delta(A_i(s))} \left( \lambda r_i(s, y_{i,s}, x_{-i,s}) + (1 - \lambda) \sum_{s' \in S} q(s' | s, y_{i,s}, x_{-i,s}) v_{i,s'} \right).$$

The set-valued function  $\varphi$  has convex and non-empty values and its graph is closed, so that by Kakutani's fixed point theorem it has a fixed point. It turns out that every fixed point of  $\varphi$  defines a  $\lambda$ -discounted equilibrium in stationary strategies.  $\square$

This result has been extended to general state and action spaces by various authors. These results assume a strong continuity on either the payoff function or the transition function, see, e.g., [39,44,62].

As in the case of zero-sum games, a dynamic programming argument shows that under a strong continuity assumption on the payoff function or on the transitions a  $T$ -stage equilibrium exists.

Regarding the existence of the limsup equilibrium and uniform equilibrium little is known. The most significant result in this direction is Vieille [77,78], who proved that every two-player stochastic game with finite state and action spaces has a uniform  $\varepsilon$ -equilibrium for every  $\varepsilon > 0$ . This result has been proved for other classes of stochastic games, see, e.g., [6,24,25,61,63,65,76]. Several influential works in this area are [22,32,71,79]. Most of the papers mentioned above rely on the vanishing discount factor approach, which constructs a uniform  $\varepsilon$ -equilibrium by studying a sequence of  $\lambda$ -discounted equilibria as the discount factor goes to 0.

For games with general state and action spaces, a limsup equilibrium exists under an ergodicity assumption on the transitions, see e.g. Nowak [44], Remark 4 and Jaskiewicz and Nowak [30].

A game has perfect information if there are no simultaneous moves, and both players observe past play. Existence of equilibrium in this case was proven by Mertens [37] in a very general setup.

### Correlated Equilibrium

The notion of correlated equilibrium was introduced by Aumann [8,9], see also Forges [▶ Correlated Equilibria and Communication in Games](#). A *correlated equilibrium* is an equilibrium of an extended game, in which each player receives at the outset of the game a private signal such that the vector of signals is chosen according to a known joint probability distribution. In repeated interactions, such as in stochastic games, there are two natural notions of correlated equilibria: (a) each player receives one signal at the outset of the game (*normal-form correlated equilibrium*); (b) each player receives a signal at each stage (*extensive-form correlated equilibrium*). It follows from Forges [26] that when the state and action sets are finite, the set of all correlated  $T$ -stage equilibrium payoffs (either normal-form or extensive-form) is a polytope.

Nowak and Raghavan [47] proved the existence of an extensive-form correlated discounted equilibrium under weak conditions on the state and action spaces. In their construction, the strategies of the players are stationary, and so is the distribution according to which the signals are chosen after each history; both depend only on the current state, rather than on the whole past play. Roughly, their approach is to apply Kakutani's fixed point theorem to the set-valued function that assigns to each game  $G_s^\lambda(v)$  the set of all correlated equilibrium payoffs in this game, which is convex and compact.

Solan and Vieille [66] proved the existence of an extensive-form correlated uniform equilibrium payoff when the

state and action spaces are finite. Their approach is to let each player play his uniform optimal strategy in a zero-sum game in which all other players try to minimize his payoff. Existence of a normal-form correlated equilibrium was proved for the class of absorbing games [68].

Solan [64] characterized the set of extensive-form correlated equilibrium payoffs for general state and action spaces and a general evaluation on the stage payoffs, and provided a sufficient condition that ensures that the set of normal-form correlated equilibrium payoffs coincides with the set of extensive-form correlated equilibrium payoffs.

### Imperfect Monitoring

So far it has been assumed that at each stage the players know the past play. There are cases in which this assumption is too strong; in some cases players do not know the complete description of the current state (Examples 3 and 4), and in others players do not fully observe the actions of all other players (Examples 2, 3 and 4). For a most general description of stochastic games see Mertens, see Chapter IV in Sorin and Zamir [40] and Coulomb [17].

In the study of the discounted equilibrium, the  $T$ -stage equilibrium or the limsup equilibrium, one may consider the game as a one-shot game: The players simultaneously choose strategies, and the payoff is either the discounted payoff, the  $T$ -stage payoff or the limsup payoff. If the strategy spaces of the players are compact (e. g., if the state and action spaces are finite), and if the payoff is upper-semicontinuous in each player's strategy, keeping the strategies of the other players fixed, then an equilibrium exists. This approach can be used successfully for the discounted equilibrium or the  $T$ -stage equilibrium under weak conditions (see, e. g., [4]), and may be used for the limsup equilibrium under a proper ergodicity condition.

Whenever there exists an equilibrium in stationary strategies (e. g., a discounted equilibrium in games with finitely many states and actions) the only information that players need in order to follow the equilibrium strategies is the current state. In particular, they need not observe past actions of the other players. As we now show, in the “Big Match” (Example 10) the limsup value and the uniform value may fail to exist when each player does not observe the past actions of the other player.

*Example 13 (Example 10: Continued.)* Assume that no player observes the actions of the other player, and assume that the initial state is  $s_2$ . Player 2 can still guarantee  $\frac{1}{2}$  in the limsup evaluation by playing the stationary strategy  $[\frac{1}{2}(L), \frac{1}{2}(R)]$ . One can show that for every strategy of player 2, player 1 has a reply such that the limsup payoff is

at least  $\frac{1}{2}$ . In other words,  $\inf_{\sigma_2} \sup_{\sigma_1} \gamma^\infty(s_2, \sigma_1, \sigma_2) = \frac{1}{2}$ . We now argue that  $\sup_{\sigma_1} \inf_{\sigma_2} \gamma^\infty(s_2, \sigma_1, \sigma_2) = 0$ . Indeed, fix a strategy  $\sigma_1$  for player 1, and  $\varepsilon > 0$ . Let  $\theta$  be sufficiently large such that the probability that under  $\sigma_1$  player 1 plays  $B$  for the first time after stage  $t$  is at most  $\varepsilon$ . Observe that as  $t$  increases, the probability that player 1 plays  $B$  for the first time after stage  $t$  decreases to 0, so that such a  $\theta$  exists. Consider the following strategy  $\sigma_2$  of player 2: Play  $R$  up to stage  $\theta$ , and play  $L$  from stage  $t + 1$  and on. By the definition of  $\theta$ , either player 1 plays  $B$  before or at stage  $\theta$ , and then the game moves to  $s_0$ , and the payoff is 0 at each stage thereafter, or player 1 plays  $T$  at each stage, and then the stage payoff after stage  $\theta$  is 0, or, with probability less than  $\varepsilon$ , player 1 plays  $B$  for the first time after stage  $\theta$ , the play moves to  $s_1$ , and the payoff is 1 thereafter. Thus, the limsup payoff is at most  $\varepsilon$ . A similar analysis shows that

$$\inf_{\sigma_2} \sup_{\sigma_1} \gamma^\lambda(s_2, \sigma_1, \sigma_2) = \frac{1}{2},$$

$$\sup_{\sigma_1} \inf_{\sigma_2} \lim_{\lambda \rightarrow 0} \gamma^\lambda(s_2, \sigma_1, \sigma_2) = 0,$$

so that the uniform value does not exist as well.

This example shows that in general the limsup value and the uniform value need not exist when the players do not observe past play. Though in general the value (and therefore also an equilibrium) need not exist, in many classes of stochastic games the value and an equilibrium do exist, even in the presence of imperfect monitoring.

Rosenberg et al. [55] and Renault [54] showed that the uniform value exists in the one player setup (Markov Decision Problem), in which the player receives partial information regarding the current state. Thus, a single decision maker who faces a dynamic situation and does not fully observe the state of the environment can play in such a way that guarantees high payoff, provided the interaction is sufficiently long or the discount factor is sufficiently low.

Altman et al. [5,6] and Flesch et al. [24] studied stochastic games in which each player has a “private” state, which only he can observe, and the state of the world is composed of the vector of private states. Altman et al. [5,6] studied the situation in which players do not observe the actions of the other players, and Flesch et al. [24] studied the situation in which players do observe each others payoffs. Such games arise naturally in wireless communication (see [5]); take for example several mobiles who periodically send information to a base station. The private state of a mobile may depend, e. g., on its exact physical environment, and it determines the power attenuation between the mobile and the base station. The throughput (the amount of bits per second) that a mobile can send to the base station depends on the power attenuations of



all the mobiles. Finally, the stage payoff is the stage power consumption.

Rosenberg et al. [57] studied the extreme case of two player zero-sum games in which the players observe neither the current state nor the action of the other player, and proved that the uniform value does exist in two classes of games, which capture the behavior of certain communication protocols. Classes of games in which the actions are observed but the state is not observed were studied, e. g., by Sorin [69,70], Sorin and Zamir [74], Krausz and Rieder [33], Flesch et al. [23], Rosenberg et al. [56], Renault [52,53]. For additional results, see [16,73].

### Algorithms

There are two kinds of algorithms: Those that terminate in a finite number of steps, and those that iterate and approximate solutions. Both kinds of algorithms were devised to calculate the value and optimal strategies (or equilibria) in stochastic games.

It is well known that the value of a two-player zero-sum matrix game and optimal strategies for the two players can be calculated efficiently using a linear program. Equilibria in two-player non-zero-sum games can be calculated by the Lemke–Howson algorithm, which is usually efficient, however, its worst running time is exponential in the number of pure strategies of the players [59]. Unfortunately, to date there are no efficient algorithms to calculate either the value in zero-sum stochastic games, or equilibria in non-zero-sum games. Moreover, in Example 9 the discounted value may be irrational for rational discount factors, even though the data of the game (payoffs and transitions) are rational, so it is not clear whether linear programming methods can be used to calculate the value of a stochastic game. Nevertheless, linear programming methods were used to calculate the discounted and uniform value of several classes of stochastic games, see [20,50]. Other methods that were used to calculate the value or equilibria in discounted stochastic games include fictitious play [80], value iterates, policy improvement, and general methods to find the maximum of a function (see [20,51]), a homotopy method [29], and algorithms to solve sentences in formal logic [15,67].

### Additional and Future Directions

The research on stochastic games extends to additional directions than those mentioned in earlier sections. We mention a few here. Approximation of games with infinite state and action spaces by finite games was discussed by Whitt [81], and further developed by Nowak [43]. Stochastic games in continuous time have also been studied, as

well as hybrid models that include both discrete and continuous aspects, see, e. g., [2,10].

Among the many directions of future research in this area, we will mention here but a few. One challenging question is the existence of a uniform equilibrium and a limsup equilibrium in multi-player stochastic games with finite state and action spaces. Another is the development of efficient algorithms that calculate the value of two-player zero-sum games. A third direction concerns the identification of applications that can be recast in the framework of stochastic games, and that can be successfully analyzed using the theoretical tools that the literature developed. Another problem that is of interest is the characterization of approachable and excludable sets in stochastic games with vector payoffs (see [12] for the presentation of matrix games with vector payoffs, and [41] for partial results regarding this problem).

### Acknowledgments

I thank Eitan Altman, János Flesch, Yuval Heller, Jean-Jacques Herings, Ayala Mashiach-Yakovi, Andrzej Nowak, Ronald Peeters, T.E.S. Raghavan, Jérôme Renault, Nahum Shimkin, Robert Simon, Sylvain Sorin, William Sudderth, and Frank Thuijmsman, for their comments on an earlier version of the entry.

### Bibliography

#### Primary Literature

- Altman E (2005) Applications of dynamic games in queues. *Adv Dyn Games* 7:309–342
- Altman E, Gaitsgory VA (1995) A hybrid (differential-stochastic) zero-sum game with fast stochastic part. *Ann Int Soc Dyn Games* 3:47–59
- Altman E, Solan E (2007) Games with constraints with networking applications. Preprint
- Altman E, Avrachenkov K, Marquez R, Miller G (2005) Zero-sum constrained stochastic games with independent state processes. *Math Methods Oper Res* 62:375–386
- Altman E, Avrachenkov K, Bonneau N, Debbah M, El-Azouzi R, Sadoc Menasche D (2008) Constrained cost-coupled stochastic games with independent state processes. *Oper Res Lett* 36:160–164
- Amir R (1996) Continuous stochastic games of capital accumulation with convex transitions. *Games Econ Behav* 15:111–131
- Aumann RJ (1974) Subjectivity and correlation in randomized strategies. *J Math Econ* 1:67–96
- Aumann RJ (1987) Correlated equilibrium as an expression of bayesian rationality. *Econometrica* 55:1–18
- Başar T, Olsder GJ (1995) *Dynamic noncooperative game theory*. Academic Press, New York
- Bewley T, Kohlberg E (1976) The asymptotic theory of stochastic games. *Math Oper Res* 1:197–208

12. Blackwell D (1956) An analog of the minimax theorem for vector payoffs. *Pac J Math* 6:1–8
13. Blackwell D, Ferguson TS (1968) The big match. *Ann Math Stat* 39:159–163
14. Chari V, Kehoe P (1990) Sustainable plans. *J Political Econ* 98:783–802
15. Chatterjee K, Majumdar R, Henzinger TA (2008) Stochastic limit-average games are in EXPTIME. *Int J Game Theory* 37:219–234
16. Coulomb JM (2003) Absorbing games with a signalling structure. In: Neyman A, Sorin S (eds) *Stochastic games and applications*. NATO Science Series. Kluwer, Dordrecht, pp 335–355
17. Coulomb JM (2003) Games with a recursive structure. In: Neyman A, Sorin S (eds) *Stochastic games and applications*. NATO Science Series. Kluwer, Dordrecht, pp 427–442
18. Dutta P, Sundaram RK (1992) Markovian equilibrium in a class of stochastic games: Existence theorems for discounted and undiscounted models. *Econ Theory* 2:197–214
19. Dutta P, Sundaram RK (1993) The tragedy of the commons? *Econ Theory* 3:413–426
20. Filar JA, Vrieze K (1996) *Competitive Markov decision processes*. Springer
21. Fink AM (1964) Equilibrium in a stochastic  $n$ -person game. *J Sci Hiroshima Univ* 28:89–93
22. Flesch J, Thuijsman F, Vrieze K (1997) Cyclic Markov equilibria in stochastic games. *Int J Game Th* 26:303–314
23. Flesch J, Thuijsman F, Vrieze OJ (2003) Stochastic games with non-observable actions. *Math Meth Oper Res* 58:459–475
24. Flesch J, Schoenmakers G, Vrieze K (2008) Stochastic games on a product state space. *Math Oper Res* 33:403–420
25. Flesch J, Thuijsman F, Vrieze OJ (2007) Stochastic games with additive transitions. *Europ J Oper Res* 179:483–497
26. Forges F (1990) Universal mechanisms. *Econometrica* 58:1341–1364
27. Fortnow L, Kimmel P (1998) Beating a finite automaton in the big match. In: *Proceedings of the 7th conference on theoretical aspects of rationality and knowledge*. Morgan Kaufmann, San Francisco, pp 225–234
28. Gillette D (1957) *Stochastic games with zero stop probabilities, contributions to the theory of games*, vol 3. Princeton University Press, Princeton
29. Herings JJP, Peeters RJAP (2004) Stationary equilibria in stochastic games: Structure, selection, and computation. *J Econ Theory* 118:32–60
30. Jaskiewicz Z, Nowak AS (2006) Zero-sum ergodic stochastic games with Feller transition probabilities. *SIAM J Control Optim* 45:773–789
31. Kakutani S (1941) A generalization of Brouwer's fixed point theorem. *Duke Math J* 8:457–459
32. Kohlberg E (1974) Repeated games with absorbing states. *Ann Stat* 2:724–738
33. Krausz A, Rieder U (1997) Markov games with incomplete information. *Math Meth Oper Res* 46:263–279
34. Levhari D, Mirman L (1980) The great fish war: An example using a dynamic Cournot–Nash solution. *Bell J Econ* 11(1):322–334
35. Maitra A, Sudderth W (1998) Finitely additive stochastic games with Borel measurable payoffs. *Int J Game Theory* 27:257–267
36. Martin DA (1998) The determinacy of Blackwell games. *J Symb Logic* 63:1565–1581
37. Mertens JF (1987) Repeated games. In: *Proceedings of the international congress of mathematicians*, American Mathematical Society, Berkeley, California, pp 1528–1577
38. Mertens JF, Neyman A (1981) Stochastic games. *Int J Game Th* 10:53–66
39. Mertens JF, Parthasarathy T (1987) Equilibria for discounted stochastic games, CORE Discussion Paper No. 8750. (Also published in *Stochastic Games and Applications*, Neyman A, Sorin S (eds), NATO Science Series, Kluwer, 131–172)
40. Mertens JF, Sorin S, Zamir S (1994) Repeated games, CORE Discussion Paper 9420-9422
41. Milman E (2006) Approachable sets of vector payoffs in stochastic games. *Games Econ Behav* 56:135–147
42. Neyman A, Sorin S (2003) *Stochastic games and applications*. NATO Science Series. Kluwer
43. Nowak AS (1985) Existence of equilibrium stationary strategies in discounted noncooperative stochastic games with uncountable state space. *J Optim Theory Appl* 45:591–620
44. Nowak AS (2003)  $N$ -person stochastic games: Extensions of the finite state space case and correlation. In: Neyman A, Sorin S (eds) *Stochastic games and applications*. NATO Science Series. Kluwer, Dordrecht, pp 93–106
45. Nowak AS (2003) On a new class of nonzero-sum discounted stochastic games having stationary Nash equilibrium points. *Int J Game Theory* 32:121–132
46. Nowak AS (2003) Zero-sum stochastic games with Borel state spaces. In: Neyman A, Sorin S (eds) *Stochastic games and applications*. NATO Science Series. Kluwer, Dordrecht, pp 77–91
47. Nowak AS, Raghavan TES (1991) Existence of stationary correlated equilibria with symmetric information for discounted stochastic games. *Math Oper Res* 17:519–526
48. Phelan C, Stacchetti E (2001) Sequential equilibria in a Ramsey tax model. *Econometrica* 69:1491–1518
49. Puterman ML (1994) *Markov decision processes: Discrete stochastic dynamic programming*. Wiley, Hoboken
50. Raghavan TES, Syed Z (2002) Computing stationary Nash equilibria of undiscounted single-controller stochastic games. *Math Oper Res* 27:384–400
51. Raghavan TES, Syed Z (2003) A policy improvement type algorithm for solving zero-sum two-person stochastic games of perfect information. *Math Program Ser A* 95:513–532
52. Renault J (2006) The value of Markov chain games with lack of information on one side. *Math Oper Res* 31:490–512
53. Renault J (2007) The value of repeated games with an informed controller. Preprint
54. Renault J (2007) Uniform value in dynamic programming. Preprint
55. Rosenberg D, Solan E, Vieille N (2002) Blackwell optimality in Markov decision processes with partial observation. *Ann Statist* 30:1178–1193
56. Rosenberg D, Solan E, Vieille N (2004) Stochastic games with a single controller and incomplete information. *SIAM J Control Optim* 43:86–110
57. Rosenberg D, Solan E, Vieille N (2006) Protocol with no acknowledgement. *Oper Res*, forthcoming
58. Sagduyu YE, Ephremides A (2003) Power control and rate adaptation as stochastic games for random access. *Proc 42nd IEEE Conf Decis Control* 4:4202–4207
59. Savani R, von Stengel B (2004) Exponentially many steps for finding a Nash equilibrium in a bimatrix game. *Proc 45th Ann IEEE Symp Found Comput Sci* 2004:258–267

60. Shapley LS (1953) Stochastic games. *Proc Nat Acad Sci USA* 39:1095–1100
61. Simon RS (2003) The structure of non-zero-sum stochastic games. *Adv Appl Math* 38:1–26
62. Solan E (1998) Discounted stochastic games. *Math Oper Res* 23:1010–1021
63. Solan E (1999) Three-person absorbing games. *Math Oper Res* 24:669–698
64. Solan E (2001) Characterization of correlated equilibria in stochastic games. *Int J Game Theory* 30:259–277
65. Solan E, Vieille N (2001) Quitting games. *Math Oper Res* 26:265–285
66. Solan E, Vieille N (2002) Correlated equilibrium in stochastic games. *Games Econ Behav* 38:362–399
67. Solan E, Vieille N (2007) Calculating uniform optimal strategies and equilibria in two-player stochastic games. Preprint
68. Solan E, Vohra R (2002) Correlated equilibrium payoffs and public signalling in absorbing games. *Int J Game Theory* 31:91–122
69. Sorin S (1984) Big match with lack of information on one side (part 1). *Int J Game Theory* 13:201–255
70. Sorin S (1985) Big match with lack of information on one side (part 2). *Int J Game Theory* 14:173–204
71. Sorin S (1986) Asymptotic properties of a non-zero-sum stochastic games. *Int J Game Theory* 15:101–107
72. Sorin S (2002) A first course on zero-sum repeated games. *Mathématiques et Applications*, vol 37. Springer
73. Sorin S (2003) Stochastic games with incomplete information. In: Neyman A, Sorin S (eds) *Stochastic Games and Applications*. NATO Science Series. Kluwer, Berlin, pp 375–395
74. Sorin S, Zamir S (1991) Big match with lack of information on one side (part 3). In: Raghavan TES et al (eds) *Stochastic games and related topics*. Kluwer, pp 101–112
75. Takahashi M (1962) Stochastic games with infinitely many strategies. *J Sci Hiroshima Univ Ser A-I* 26:123–134
76. Thuijsman F, Raghavan TES (1997) Perfect information stochastic games and related classes. *Int J Game Theory* 26:403–408
77. Vieille N (2000) Equilibrium in 2-person stochastic games I: A Reduction. *Israel J Math* 119:55–91
78. Vieille N (2000) Equilibrium in 2-person stochastic games II: The case of recursive games. *Israel J Math* 119:93–126
79. Vrieze OJ, Thuijsman F (1989) On equilibria in repeated games with absorbing states. *Int J Game Theory* 18:293–310
80. Vrieze OJ, Tijs SH (1982) Fictitious play applied to sequences of games and discounted stochastic games. *Int J Game Theory* 12:71–85
81. Whitt W (1980) Representation and approximation of noncooperative sequential games. *SIAM J Control Optim* 18:33–48
- Mertens JF, Sorin S, Zamir S (1994) Repeated games. CORE Discussion Paper 9420-9422
- Raghavan TES, Shapley LS (1991) Stochastic games and related topics: In honor of Professor L.S. Shapley. Springer
- Vieille N (2002) Stochastic games: Recent results. In: Aumann RJ, Hart S (eds) *Handbook of game theory with economic applications*, vol 3. Elsevier, pp 1833–1850

## Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems

HANS C. FOGEDBY<sup>1,2</sup>

<sup>1</sup> Department of Physics and Astronomy, University of Aarhus, Aarhus, Denmark

<sup>2</sup> Niels Bohr Institute, Copenhagen, Denmark

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Scaling](#)

[Conformal Invariance](#)

[Loewner Evolution](#)

[Stochastic Loewner Evolution](#)

[Results and Discussion](#)

[Future Directions](#)

[Acknowledgments](#)

[Bibliography](#)

### Glossary

**Bessel process** The Bessel process operates in  $d$  dimensions and describes the radial distance  $r$  from the origin of a particle performing Brownian motion. The random motion is governed by the Langevin equation  $dr/dt = \kappa(d-1)/2r + \xi$ , where  $\langle \xi \xi \rangle(t) = \kappa \delta(t)$ . For  $d \leq 2$  the motion is recurrent, i.e., returns to the origin; for  $d > 2$  the motion goes off to infinity. In SLE the Bessel process describes the transition between simple curves and self-intersecting curves.

**Brownian motion** Brownian motion,  $B_t$ , is the scaling limit of random walk. Brownian motion is plane-filling, has the fractal dimension  $D = 2$ , and is described by the Langevin equation  $dB_t/dt = \eta_t$ ,  $\langle \eta_t \eta_s \rangle = \delta(t-s)$ .  $B_t$  is characterized by i) the stationarity property,  $B_{t+t'} - B_t$  and  $B_{t'}$  identical in distribution and ii) the independence property,  $B_{\Delta t'}$  and  $B_{\Delta t}$  independent for  $\Delta t \neq \Delta t'$ . The correlations are given by

### Books and Reviews

- Başar T, Olsder GJ (1995) *Dynamic noncooperative game theory*. Academic
- Filar JA, Vrieze K (1996) *Competitive Markov decision processes*. Springer
- Maitra AP, Sudderth WD (1996) *Discrete gambling and stochastic games*. Springer
- Mertens JF (2002) Stochastic games. In: Aumann RJ, Hart S (eds) *Handbook of game theory with economic applications*, vol 3. Elsevier, pp 1809–1832

$\langle |B_t - B_s|^2 \rangle = |t - s|$  and  $B_t$  is distributed according to the Gaussian (normal) distribution  $P(B, t) = (2\pi t)^{-1/2} \exp(-B^2/2t)$ .

**Chordal SLE** In order to map the geometry of the growing hull to a real function by means of SLE one chooses a reference domain. Chordal SLE refers to the case where the SLE trace is grown between two boundary points, usually the origin and the point at infinity in the upper half complex plane.

**Conformal invariance** Conformal invariance or local scale invariance is a larger symmetry than scale invariance. Conformal invariance implies invariance under both a local rotation, translation, and dilatation. Conformal invariance is particularly powerful in 2D where a conformal transformation is implemented by an analytic function.

**Conformal transformation** A conformal transformation is a transformation which preserves angles but allows for rotation, dilatation, and translation. In an elastic medium picture a conformal transformation corresponds to translation, rotation, and compression (dilatation), but not shear. In 2D an analytic complex function  $w = f(z)$  from the complex  $z$  plane to the complex  $w$  plane generates a conformal transformation.

**Continuum limit** The limit where a lattice model approaches a continuum model; also called the scaling limit. The resulting continuum field theories define the universality classes of the lattice models.

**Correlation length** The correlation length measures the size of correlations. At the critical point the correlation length diverges signaling that the system becomes scale invariant.

**Critical curves** Critical curves are domain walls or cluster boundaries at the critical point. Critical curves are scale invariant and characterized by a fractal dimension.

**Exploration** Domain walls at the critical point can mathematically be generated by an exploration process where the domain wall is initiated at a boundary point and constructed in steps across the domain. In the percolation case the local step is generated by 'flipping a coin', in the Ising case by evaluating the magnetization at the tip of the 'growing' domain wall. The construction by an exploration process is essential in generating a critical curve by stochastic Loewner evolution.

**Fortuin–Kasteleyn representation (FK)** Based on a high temperature expansion one can represent the configurations in the Ising,  $O(n)$ , and Potts models by means of the Fortuin–Kasteleyn representation in terms of

random clusters. The FK transformation is essential in identifying random curves for the lattice models which then can be accessed by SLE in the scaling limit.

**Fractal dimension** Irregular objects can be characterized by a fractal dimension. The fractal dimension  $D$  is derived by covering the object with  $N(\ell)$  intervals, disks, or spheres of linear dimension  $\ell$ . By letting  $\ell \rightarrow 0$  this definition resolves the fine scale fractal structure and the scaling relation  $N(\ell) \propto (\ell)^{-D}$ , or  $D = -\lim_{\ell \rightarrow 0} \ln(N(\ell)/\ln(\ell))$ , yields the fractal dimension. Examples of deterministic fractals are for example the Cantor set, the Koch curve, the Sierpinski gasket, and plane-filling Hilbert and Peano curves. Random walk and diffusion limited aggregation (DLA) are examples of random physical fractals.

**Hulls** For a self-intersecting SLE trace the trace and enclosed regions form a so-called hull. Points in the hull cannot be connected to infinity without crossing the trace. The growing hull of a self-intersecting SLE trace eventually exhausts the half plane.

**Ising model** The Ising model originates from the theory of magnetism and plays an important role in the theory of phase transitions and in general in statistical mechanics. The model is defined on a lattice where each lattice sites is endowed with a local spin variable assuming two distinct values. The spins interact by a short range exchange interaction. Above 1D the Ising model has a second order phase transition.

**Locality** In the percolation case the domain wall is constructed by an exploration process and a local rule for assigning the next step. This locality property is specific to percolation which has a geometric phase transition. In the SLE context the locality property implies  $\kappa = 6$ , i. e., the percolation case.

**Loewner equation** The Loewner equation is the first order nonlinear differential equation for the uniformizing map  $g_t(z)$  which maps the complement of a growing hull or curve in the upper half plane back to the upper half plane. The geometrical properties of the hull is encoded in the real function  $a_t$ . The Loewner equation has the form  $dg_t/dt = 2/(g_t - a_t)$ .

**Loewner evolution** Loewner evolution refers to the parametrized map  $g_t(z)$  which uniformizes the complement of a curve or hull, say in the upper half plane.

**Loop erased random walk (LERW)** Loop erased random walk is a variant of random walk where loops formed are removed as the walk progresses. LERW has a conformally invariant scaling limit and can be generated by SLE for  $\kappa = 2$ .

**Markov property** The Markov property refers to a stochastic process without memory. A typical ex-



ample is random walk or in the scaling limit Brownian motion. The Markov property is essential for the application of SLE to domain walls in the scaling limit.

**Measures** For lattice systems one can define probability distributions according to the rules of statistical mechanics. In the scaling limit the distributions typically diverge and have to be replaced by the more abstract mathematical concept of probability measures.

**$O(n)$  models** The  $O(n)$  model is a generalization of the Ising model. At each lattice site is associated an  $n$ -component spin variable. For  $n = 1$  we recover the Ising model,  $n = 2$  corresponds to the XY model, and  $n = 3$  to the Heisenberg model.

**Peano curve** A Peano curve is a non-crossing curve which is dense in the plane, i. e., it gets arbitrarily close to every point. The Peano curve has the fractal dimension  $D = 2$ . In a SLE context a random Peano curve winds around a UST and corresponds to  $\kappa = 8$ .

**Percolation** Percolation on a lattice is constructed by filling lattice sites at random with a common probability  $p$ . At a critical concentration a spanning cluster extends across the system.

**Potts model** The Potts model is a generalization of the Ising model where the local lattice variable can assume  $q$  different values and where sites only interact when they are in the same Potts state. The Potts model has a FK cluster representation. The Ising model corresponds to  $q = 2$ ;  $q = 1$  corresponds to percolation and  $q = 0$  to the UST.

**Random walk** Random walk is an ubiquitous phenomenon in nature. In random walk on a square lattice a particle jump from site to site with a given fixed probability. Each step is independent of the past history, there is no memory, this is the so-called Markov property. The random walk path or history is plane-filling and has the fractal dimension  $D = 2$ . The mean square displacement of random walk scales with the number of steps.

**Restriction** The restriction property in a SLE context implies that the measure on a curve in a domain  $D$  conditioned not to hit a bulge  $L$  is the same as the measure in the domain  $D \setminus L$ . The restriction property holds in the case of a uniform measure and applies to self-avoiding random walk.

**Riemann's mapping theorem** Riemann's mapping theorem states that an arbitrary simply connected domain  $D$ , i. e., without holes, can be mapped to another simply connected domain  $D'$  by means of a suitable complex function  $g(z)$ , i. e.,  $g(D) = D'$ . Often the disk or half plane are used as reference domains. The map-

ping theorem does not make assumptions about the domain boundaries which can be fractal.

**Scale invariance** The scaling property refers to the case where a phenomenon is devoid of a characteristic scale or unit. Scale invariance is typically characterized by a power law behavior and critical exponents.

**Scaling limit** The limit where the lattice parameter in a lattice model approaches zero. The scaling limit is equivalent to the continuum limit.

**Self-avoiding random walk (SAW)** A self-avoiding random walk is a walk conditioned not to cross itself. SAW has been used to model polymers. SAW has a uniform probability measure and conforms in a SLE context to the restriction condition.

**Stochastic Loewner evolution (SLE)** Stochastic Loewner evolution is Loewner evolution driven by a real stochastic function  $a_t$  with a distribution given by 1D Brownian motion, i. e.,  $a_t = \sqrt{\kappa} B_t$ . SLE is governed by the stochastic equation of motion  $dg_t/dt = 2/(g_t - a_t)$ .

**Schramm's theorem** Schramm's theorem refers to Schramm's derivation of SLE for LERW. Schramm showed that the scaling limit of LERW is described by SLE for  $\kappa = 2$ . Schramm also conjectured that the random Peano curve winding around an UST is described by SLE for  $\kappa = 8$  and that percolation is described by SLE for  $\kappa = 6$ .

**Uniformizing maps** Conformal transformations which map a domain  $D$  to a standard reference domain, e. g., the half plane or the disk, are called uniformizing maps.

**Uniform spanning tree (UST)** A spanning tree is a collection of vertices and links forming a tree (no loops or cycles). A USF is a random tree picked among all spanning trees with equal probability.

## Definition of the Subject

Stochastic Loewner evolution also called Schramm Loewner evolution (abbreviated, SLE) is a rigorous tool in mathematics and statistical physics for generating and studying scale invariant or fractal random curves in two dimensions (2D). The method is based on the older deterministic Loewner evolution introduced by Karl Löwner [64], who demonstrated that an arbitrary curve not crossing itself can be generated by a real function by means of a conformal transformation. A real function defined in one spatial dimension (1D) thus encodes a curve in 2D, in itself an intriguing result. In 2000 Oded Schramm [77] extended this method and demonstrated that driving the Loewner evolution by a 1D Brownian mo-



tion, the curves in the complex plane become scale invariant; the fractal dimension turns out to be determined by the strength of the Brownian motion.

The one parameter family of scale invariant curves generated by SLE is conjectured and has in some cases been proven to represent the continuum or scaling limit of a variety of interfaces and cluster boundaries in lattice models in statistical physics, ranging from self-avoiding random walks to percolation cluster boundaries, and Ising domain walls.

SLE operates in the 2D continuum where it generates extended scale invariant objects. SLE delimits scaling universality classes by a single parameter  $\kappa$ , the strength of the 1D Brownian drive, yielding the fractal dimension  $D$  of the scale invariant shapes according to the relation  $D = 1 + \kappa/8$ . Moreover, SLE provides the geometrical aspects of conformal field theory (CFT). The central charge  $c$ , delimiting scaling universality classes in CFT, is thus related to  $\kappa$  by means of the expression  $c = (3\kappa - 8)(6 - \kappa)/2\kappa$ .

Stochastic Loewner evolution derives its importance from the fact that it addresses the issue of extended random fractal shapes in 2D by direct analysis in the continuum. It thus supplements and extends earlier lattice results and also allows for the determination of new scaling exponents. From the point of view of conformal field theory based on the concept of a local field, operator expansions, and correlations, the geometrical approach afforded by SLE, directly addressing conformally invariant random shapes in the continuum, represents a novel point of view of maybe far reaching consequences; so far only explored in two dimensions.

## Introduction

The field of SLE has mainly been driven by mathematicians presenting their results in long and difficult papers. There are, however, presently several excellent reviews of SLE both addressing the theoretical physics community [11,13,25,27,39,43] and the mathematical community [56,85], see also a complete biography up to 2003 [39]. The purpose of this article is to present a heuristic and simple discussion of some of the key aspects of SLE, for details and topics left out we refer the reader to the reviews mentioned above. However, in order to provide the necessary background and set the stage for SLE we begin with some general remarks on scaling in statistical physics.

## General Remarks

In statistical physics we study macroscopic systems composed of many interacting components. In the limit

of many degrees of freedom the macroscopic behavior roughly falls in two categories. In the most common case the macroscopic behavior is deterministic and governed by phenomenological theories like for example thermodynamics and hydrodynamics operating entirely on a macroscopic level. This behavior is basically a result of the law of large numbers, permitting an effective coarse graining and yielding for example a macroscopic density or velocity field [29]. In the other case, the macroscopic behavior is dominated by fluctuations and shows a random behavior [29,73]. Typical cases are random walk and equilibrium systems tuned to the critical point of a second order phase transition. Other random cases are for example self-organized critical systems purported to model earth quake dynamics, flicker noise and turbulence in fluids [5,40].

The distinction between the deterministic and random cases of macroscopic behavior is illustrated by the simple example of a biased random walk described by the Langevin equation  $dx(t)/dt = v + \xi(t)$ ,  $\langle \xi(t)\xi(0) \rangle \propto \delta(t)$ . Here  $x(t)$  is a macroscopic variable sampling the statistically independent microscopic steps  $\xi(t)$ ; the velocity  $v$  is an imposed drift or bias. Averaging over the steps we obtain for the deviation of  $x$ ,  $R = [\langle x^2 \rangle]^{1/2} = [(vt)^2 + t]^{1/2}$ . For large  $t$  the deviation  $R \sim \langle x \rangle = vt$  and the mean value or deterministic part dominates the behavior, the fluctuational or random part  $R \sim t^{1/2}$  being subdominant. Fine tuning the random walk to vanishing bias  $v = 0$  we have  $\langle x \rangle = 0$  and  $R \sim t^{1/2}$ , i. e., the random fluctuations control the phenomenon.

The study of complexity encompasses a broader field than statistical physics and is concerned with the emergence of universal properties on a mesoscopic or macroscopic scale in large interacting systems. For example particle systems, networks in biology and sociology, cellular automata, etc. The class of complex systems generally falls in the category of random systems. The emergent properties are a result of many interacting agents or degrees of freedom and can in general not be directly inferred from the microscopic details. A major issue is thus the understanding of generic emerging properties of complex systems [45,70,83]. Here, however, the methods of statistical physics is an indispensable tool in the study of complexity.

The evolution of statistical physics, a branch of theoretical physics, has occurred in steps and is driven both by the introduction of new concepts and the concurrent development of new mathematical methods and techniques. A well-known case is the long standing problem of second order phase transitions or critical phenomena which gave way to a deeper understanding in the sixties and seventies and spun off the renormalization group techniques

for the determination of critical exponents and universality classes [20,24,65,81,86].

### Scaling Ideas

This brings us to the fundamental scaling ideas and techniques developed particularly in the context of critical phenomena in equilibrium systems and which now pervade a good part of theoretical physics and, moreover, play an important role in the analysis of complex systems in physical sciences [20,24,29]. Scaling is synonymous with no scale in the sense that a system exhibiting scale invariance is characterized by the absence of any particular scale or unit. Scaling occurs both in the space and/or time behavior and is typically characterized by power law dependencies controlled by scaling exponents.

A classical case is random walk discussed above, characterized by the Langevin equation  $dx/dt = \xi(t)$ ,  $\langle \xi \xi \rangle \sim \delta(t)$  [73]. Here the mean square displacement scales like  $\langle x^2 \rangle(t) \sim t^{2H}$ , where  $H$  is the Hurst scaling exponent; for random walk  $H = 1/2$  [32,68]. Correspondingly, the power spectrum  $P(\omega) = |x_\omega|^2$ ,  $x_\omega = \int dt x(t) \exp(i\omega t)$ , scales like  $P(\omega) \sim \omega^{-1-2H}$ , i. e.,  $P(\omega) \sim \omega^{-2}$  for random walk; we note that the underlying reason for the universal scaling behavior of random walk is the central limit theorem [37,73].

### Scaling in Equilibrium

Scaling ideas and associated techniques first came to the forefront in statistical physics in the context of second order phase transitions or critical phenomena [20,24,29,65]. More specifically, consider the usual Ising model defined on a lattice with a local spin degrees of freedom,  $\sigma_i = \pm 1$  at site  $i$ , subject to a short range interaction  $J$  favoring spin alignment. The Ising Hamiltonian has the form  $H = -J \sum_{\langle ij \rangle} \sigma_i \sigma_j$ , where  $\langle ij \rangle$  indicates nearest neighbor sites. The thermodynamic phases are characterized by the order parameter  $m = \langle \sigma_i \rangle$ . Above one dimension the Ising model exhibits a second order phase transition at a finite critical temperature  $T_c$ . Above  $T_c$  the model is in a disordered paramagnetic state with  $m = 0$  and only microscopic domain of ordered spins. Below  $T_c$  the model favors a ferromagnetic state with long range order and macroscopic domains of ordered spins, corresponding to  $m \neq 0$ ; at  $T = 0$  the model assumes the ferromagnetic ground state configuration with totally aligned spins. Regarding the spatial organization, the size of the domains of ordered spins is characterized by the correlation length  $\xi(T)$ . As we approach the critical point at  $T_c$  the order parameter vanishes  $m \propto |T - T_c|^\beta$  with critical exponent  $\beta$ , but more significantly, the correlation length  $\xi(T)$  diverges

like  $\xi(T) \sim |T - T_c|^{-\nu}$  with scaling exponent  $\nu$ . This indicates that the system is scale invariant at  $T_c$ . Regarding the domains of ordered spins, the system is spatially self-similar on scales from the microscopic lattice distance to the system size; the system size diverging in the thermodynamic limit. The scaling behavior at the critical point  $T_c$  is an emergent property in the sense that the scaling exponents  $\beta$  and  $\nu$  do not depend on microscopic details like the type of lattice, strength of interaction, etc., but only on the dimension of the system and the symmetry of the order parameter [72].

The diverging correlation length at the critical point is the central observation which in the 60-ties and 70-ties gave rise to a detailed understanding of critical phenomena, beginning with the coarse graining block scheme proposed by Kadanoff [41] and culminating with the development and application of field theoretical renormalization group techniques by Wilson and others [20,24,29,65,72,86].

For a diverging correlation length much larger than the lattice distance the local spin  $\sigma_i$  can be replaced by a coarse-grained local field  $\phi(r)$  and the Ising Hamiltonian  $H$  by the Ginzburg–Landau functional  $F = \int d^d r [(\nabla \phi)^2 + R\phi^2 + U\phi^4]$ , where the ‘mass’ term  $R \sim |T - T_c|$ . Consequently, the universality class of Ising-type models is described by a scalar field theory. The renormalization group techniques basically quantify the Kadanoff block construction in momentum space and extract scaling properties in an expansion about the upper critical dimension  $d = 4$ . To leading order in  $4 - d$  one obtains  $\beta = 1/2 - (4 - d)/6$  and  $\nu = 1/2 + (4 - d)/12$ . Alternatively, keeping the correlation length  $\xi$  fixed and letting the lattice distance approach zero, we obtain at  $T_c$  the so-called scaling limit or continuum limit of the Ising model. The Ising spin  $\sigma_i$  becomes a local field  $\phi(r)$  and the weight of a configuration is determined by  $\exp(-F)$ . Note that in order to implement the scaling limit we must be at  $T_c$ . The two scenarios of a growing correlation length for fixed lattice distance implementing the Kadanoff construction and a fixed correlation length for a vanishing lattice distance yielding a continuum field theory are related by an overall scale transformation [86].

It is generally assumed that the global or nonlocal scale invariance at the critical point in the continuum limit can be extended to a local scale invariance including translation and rotation, that is an angle-preserving conformal transformation. This follows heuristically from an a local implementation of the Kadanoff coarse-graining block construction and applies to lattice models with short range interactions and discrete translational and rotational invariance [22,24]. The resulting continuum the-

ories then fall in the category of conformal field theories (CFT).

In 2D the group of conformal transformations is particularly rich since it corresponds to the class of analytical functions  $w = f(z)$ , mapping the complex plane  $z$  to the complex plane  $w$ . The infinite group structure imposes sufficient constraints on the structure of conformal field theories in 2D that the scaling form of correlations, e. g.,  $\langle \phi \phi \rangle(r)$ , and in particular the critical exponents can be determined explicitly [23,24]. One finds  $\beta = 1/8$  and  $\nu = 1$  for the order parameter and correlation length exponents, respectively, in accordance with lattice theory results [14]. Here we also mention the Coulomb gas method for the determination of critical exponents [71].

It is a common feature of both renormalization group calculations based on an expansion about a critical dimension and conformal field theory in 2D that the local field  $\phi(r)$  and its correlations are the basic building blocks and that the critical properties are encoded in their scaling form, yielding critical exponents, scaling laws, scaling functions, etc. Notwithstanding the fact that the seminal Kadanoff construction [41] was based on a geometrical picture corresponding to coarse-graining the degrees of freedom over larger and larger scales, keeping track of ordered domains on all scales, the actual geometry of critical phenomena such as the scaling properties of critical clusters was not well-understood and seemed inaccessible in the continuum limit within the context of local field theories. Whereas it is not difficult to generate critical domain walls, interfaces, and clusters for lattice models with appropriate boundary conditions by means of standard Monte Carlo simulation techniques, the continuum or scaling limit of critical shapes appeared until recently, with a few exceptions, beyond present techniques.

### Stochastic Loewner Evolution

Here stochastic Loewner evolution (SLE) represents a new insight in 2D critical phenomena with respect to a deeper understanding of scale invariant curves, clusters, and shapes. Also, there appears to be deep connections between SLE and conformal field theory.

SLE is an ingenious way of generating critical curves and shapes in the 2D continuum using conformal transformations. Let us mention a characteristic example. Consider an Ising model on a lattice in a chosen domain. Imposing spin up on a continuous part of the domain boundary and spin down on the remaining part of the boundary, it follows that a specific domain wall or interface will connect the two points on the boundary where a bond is broken.

At low temperature the bond energies dominate and the free energy is lowest for a straight domain wall with few kinks. As we approach the critical point entropy or fluctuations come strongly into play and the domain wall meanders balancing energy and entropy. At the critical point the system becomes scale invariant with a diverging correlation length and likewise the domain wall becomes scale invariant, i. e., it has kinks on all scales larger than the lattice distance. In the continuum limit the Ising domain wall becomes a random fractal curve with a particular fractal dimension. Here SLE provides a direct analytical method in the continuum to generate such a random curve and, moreover, provides the fractal dimension in terms of the strength of the 1D random walk driving the SLE evolution.

### Outline

The outline of the present article is as follows. In Sect. “[Scaling](#)” on scaling we introduce some of the basic models and concepts necessary for a discussion of SLE: A) Random walk, B) Percolation, C) Ising model, D) Critical curves and exploration, and E) Distributions, Markov properties, and measures. In Sect. “[Conformal Invariance](#)” we turn to the essential ingredient in SLE, namely, conformal invariance: A) Conformal maps and B) Measures and conformal invariance. Section “[Loewner Evolution](#)” is devoted to deterministic or classical Loewner evolution: A) Growing stick, B) Loewner equation and C) Exact solutions. In Sect. “[Stochastic Loewner Evolution](#)”, constituting the core part of this article, we discuss stochastic Loewner evolution: A) Schramm’s theorem, B) SLE properties, C) Curves, hulls, and the Bessel process and D) Fractal dimension. Section “[Results and Discussion](#)” is devoted to results and discussions: A) Phase transitions, locality, restriction, and duality, B) Loop erased random walk, C) Self-avoiding random walk, D) Percolation, E) Ising model and  $O(n)$  models, F) SLE and conformal field theory, G) Application to 2D turbulence, H) Application to 2D spin glass and I) Further remarks. Finally, in Sect. “[Future Directions](#)” we discuss future directions of the field. The bibliography includes books, general reviews, and more technical papers.

### Scaling

Stochastic Loewner evolution, has been applied to a host of lattice models proved or conjectured to possess scaling limits. However, for the present purpose we will focus on three lattice models: Random walk, Percolation, and the Ising model.

## Random Walk

Random walk is a simple and much studied random process [4,32,73]. Consider an unbiased random walk in the plane composed of  $N$  steps, where the  $i$ th step,  $\vec{\eta}_i$ , is random, isotropic and uncorrelated, i. e.,  $\langle \vec{\eta}_i \rangle = 0$  and  $\langle \eta_i^\alpha \eta_j^\beta \rangle \propto \delta_{\alpha\beta} \delta_{ij}$ . For the end-to-end distance we have  $\vec{x} = \sum_{i=1}^N \vec{\eta}_i$  and for the size  $R = [\langle \vec{x}^2 \rangle]^{1/2} \sim N^{1/2}$ . Assuming one step per unit time,  $N \propto t$ , we obtain  $R \sim t^{1/2}$ , characteristic of diffusive motion. Introducing the fractal dimension  $D$  by the usual box counting procedure [32,68]

$$N(R) \sim R^D, \quad (1)$$

where  $N$  is the number of boxes and  $R$  the size of the object, and covering the random walk we readily infer  $D = 2$ , i. e., the self-crossing random walk is plane-filling modulo the lattice distance. Introducing the scaling exponent  $\nu$  according to  $R \sim N^\nu$  we have for random walk  $\nu = 1/2$ ; note that  $D = 1/\nu$ .

The scaling limit of unbiased random walk is Brownian motion (BM) [4] and is obtained by scaling the step size  $\eta$  down and the number of steps  $N$  up in such a manner that the size  $R \sim N^{1/2} \eta$  stays constant. The resulting BM path is a continuous non-differentiable random curve with fractal dimension  $D = 2$ . The BM path is plane-filling and recurrent in 2D, i. e., it returns to a given point with probability one. Focusing on one of the independent cartesian components 1D BM,  $B_t$ , is also described by the Langevin equation

$$\frac{dB_t}{dt} = \eta_t, \quad \langle \eta_t \eta_s \rangle = \delta(t-s), \quad (2)$$

where  $\eta_t$  is uncorrelated Gaussian white noise with a flat power spectrum. Integrating Eq. (2)  $B_t$  samples the step  $\eta_t$  and we find, assuming  $B_0 = 0$ ,  $B_t = \int_0^t \eta_{t'} dt'$  from which we directly infer the fundamental properties of BM, namely, independence and stationarity,

$$B_{T+\Delta T} - B_T \approx B_{\Delta T}, \quad (\text{stationarity}) \quad (3)$$

$$B_{\Delta T}, B_{\Delta T'} \text{ indep. for } \Delta T \neq \Delta T', \quad (\text{independence}) \quad (4)$$

where  $\approx$  indicates identical distributions. Moreover, the correlations are given by

$$\langle |B_t - B_s|^2 \rangle = |t - s| \quad (5)$$

and  $B_t$  is distributed according to the normal (Gaussian) distribution  $P(B, t) = (2\pi t)^{-1/2} \exp(-B^2/2t)$ .

Whereas 1D BM drives SLE, 2D BM is not itself generated by SLE since the path is self-crossing on all scales; by construction SLE is limited to the generation of non-crossing random curves. However, variations of BM have played an important role in the development of SLE. We mention here the scaling limit of loop erased random walk (LERW) and self-avoiding random walk (SAW), to be discussed later.

## Percolation

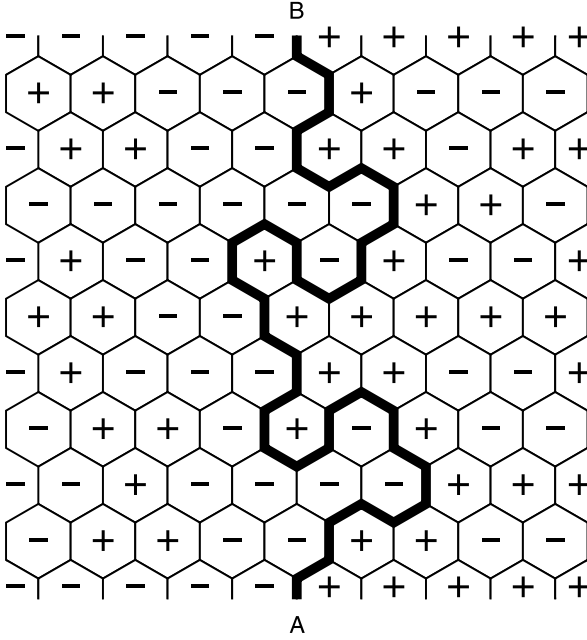
The phenomenon of percolation is relevant in the context of clustering, diffusion, fractals, phase transitions and disordered systems. As a result, percolation theory provides a theoretical and statistical background to many physical and natural sciences [82].

Percolation is the simplest lattice model exhibiting a geometrical phase transition. The site percolation model is constructed by occupying sites on a lattice with a given common probability  $p$ . Let an occupied site be denoted 'plus' and an 'empty' site denoted 'minus'. For  $p$  close to zero the sites are mainly unoccupied and the lattice is 'minus'. For  $p$  close to one the sites are predominantly occupied and the lattice is 'plus'. At a critical concentration  $p_c$ , the percolation threshold, an infinite cluster of 'plus' sites embedded in the 'minus' background extends across the lattice. In the scaling limit of vanishing lattice distance the critical cluster has a fractal boundary which can be accessed by SLE. Whereas the scaling properties of critical percolation clusters define a universality class and is independent of the lattice structure, the critical concentration  $p_c$  in general depends on the lattice. For site percolation on a triangular lattice the percolation threshold is known to be  $p_c = 1/2$ .

In Fig. 1 we depict a realization of site percolation on a triangular lattice in the upper half plane at the percolation threshold  $p_c = 1/2$ . The occupied sites are denoted 'plus', the empty sites 'minus'. By imposing appropriate boundary conditions we induce a meandering domain wall across the system from A to B. In the scaling limit the domain wall becomes a fractal non-crossing critical curve.

## Ising Model

The Ising model is probably the simplest interacting many particle system in statistical physics [20,29,81]. The model has its origin in magnetism but has become of paradigmatic importance in the context of phase transitions. The model is defined on a lattice where each lattice site  $i$  is occupied by a single degree of freedom, a spin variable  $\sigma_i$ , assuming two values,  $\sigma_i = \pm 1$ , i. e., spin up or spin down.



**Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 1**

We depict site percolation on a triangular lattice in the *upper half plane* at the percolation threshold. The critical concentration is  $p_c = 1/2$ . The occupied sites are denoted ‘plus’, the empty sites ‘minus’. The boundary conditions enforce a meandering domain wall from A to B

In the ferromagnetic case considered here the spins interact via a short range exchange interaction  $J$  favoring parallel spin alignment. The model is described by the Ising Hamiltonian

$$H = -J \sum_{\langle ij \rangle} \sigma_i \sigma_j, \quad (6)$$

where  $\langle ij \rangle$  indicates nearest neighbor spin sites  $i$  and  $j$ . The statistical weight or probability of a specific spin configuration  $\{\sigma_i\}$  is determined by the Boltzmann factor

$$P(\{\sigma_i\}) = \exp[-H/kT]/Z, \quad (7)$$

where  $T$  is the temperature and  $k$  Boltzmann’s constant. The partition function  $Z$  has the form

$$Z = \sum_{\{\sigma_i\}} \exp(-H/kT), \quad (8)$$

yielding the thermodynamic free energy  $F$  according to  $F = -kT \log Z$ . The entropy is given by  $S = -dF/dT$  and the energy follows from  $F = E - TS$ . The magnetization

or order parameter and correlations are given by

$$m = \sum_{\{\sigma_i\}} \sigma_i P(\{\sigma_i\}) \quad \text{and} \quad \langle \sigma_i \sigma_j \rangle = \sum_{\{\sigma_i\}} \sigma_i \sigma_j P(\{\sigma_i\}), \quad (9)$$

respectively.

The Ising model possesses a phase transition at a critical temperature  $T_c$  from a disordered paramagnetic phase above  $T_c$  with vanishing order parameter  $m = 0$  to a ferromagnetic ordered phase below  $T_c$  with non-vanishing order parameter  $m \neq 0$ . At the critical temperature  $T_c$  the order parameter vanishes like  $m \sim |T - T_c|^\beta$  with critical exponent  $\beta$ . The correlation function  $\langle \sigma_i \sigma_j \rangle$  monitoring the spatial organization of ordered domains behaves like

$$\langle \sigma_i \sigma_j \rangle \sim \frac{\exp[-|i - j|/\xi]}{|i - j|^\eta}, \quad (10)$$

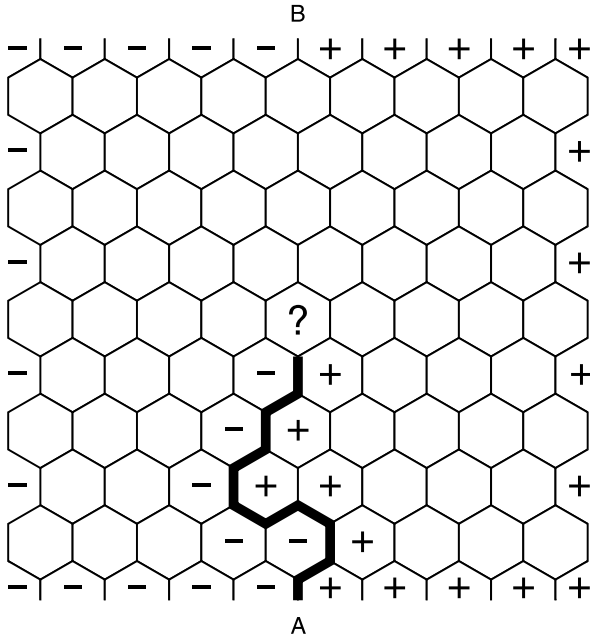
where  $|i - j|$  denotes the distance between position  $i$  and position  $j$ . The algebraic behavior is characterized by the critical exponent  $\eta$  and the correlation length  $\xi$  scales like  $\xi \sim |T - T_c|^{-\nu}$  with critical exponent  $\nu$ . At the critical point the correlation length  $\xi$  diverges and the Ising model becomes scale invariant with an algebraically decaying correlation function  $\langle \sigma_i \sigma_j \rangle \sim |i - j|^{-\eta}$ . Assigning ‘plus’ to spin up and ‘minus’ to spin down, Fig. 1 also illustrates a typical configuration of the 2D Ising model on a triangular lattice at the critical temperature, including an Ising domain wall from A to B.

### Critical Curves – Exploration

In the percolation case at the percolation threshold a critical curve is induced by fixing the boundary conditions. Occupying sites from A to B along the right hand side of the boundary and assigning empty sites along the left hand side of the boundary from A to B a critical curve will meander across the system from A to B. Imagining painting the two sides of the curve, one side is painted with occupied sites, the other side with empty sites. Typically the curve meanders on all scales but by construction does not cross itself. For later purposes the configuration is depicted in Fig. 1.

In order to make contact with SLE we observe that a critical interface in the percolation case also can be constructed by an exploration process. We initiate the curve at the boundary point A and toss a coin. In the case of ‘head’ the site or hexagon in front is chosen to be occupied and the path bends left; in the event of ‘tail’ the hexagon in front is left unoccupied and the path bends right. In this manner a critical meandering non-crossing curve is generated terminating eventually at B. The percolation growth





**Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 2**

We depict the growth process in the percolation case. The percolation threshold is at  $p_c = 1/2$ . The interface imposed by the boundary conditions originates at the boundary point A and progresses towards the boundary point B

process is depicted in Fig. 2, where we for clarity only have indicated the sites involved in the growth process. We note that since there is no interaction between the sites the path depends entirely on the local properties.

In the case of the Ising model a critical interface or domain walls at the critical point is again fixed by assigning appropriate boundary conditions with spin up along the boundary from A to B and spin down from B to A. The critical curve can be constructed in two ways: Globally or by an exploration process. In the global case we generate a spin configuration by means of a Monte Carlo simulation, i.e., perform a biased importance sampling implementing the probability distribution in Eq. (7), and identify a critical interface. Alternatively, we can generate an interface by an exploration process like in the percolation case, occupying a site  $i$  according to the weight  $(1/2)(1 + \langle \sigma_i \rangle)$ , where  $\langle \sigma_i \rangle$  is evaluated in the domain with the spins along the interface fixed, see again Fig. 2.

### Distributions – Markov Properties – Measures

In the case of random walk the number of walks of length  $L$  grows like  $\mu^L$ , where  $\mu$  is a lattice dependent number, i.e., at each step there are  $\mu$  lattice-dependent choices for

choosing a direction of the next step. Consequently, the weight or probability of a particular walk of length  $L$  is proportional to  $\mu^{-L}$ ,

$$P(L) \sim \mu^{-L}, \quad (11)$$

and all walks of a given length have the same weight. We note here the important Markov property, characteristic of random walk, which can be formulated in the following manner. Assume that the first part  $\gamma'$  of the walk has taken place and thus conditions the remaining part  $\gamma$  of the walk. In a suggestive notation the conditional distribution is given by  $P(\gamma|\gamma') = P(\gamma\gamma')/P(\gamma')$ . The Markov property implies that the conditional probability is equal to the probability of the walk  $\gamma$  in a new domain where the first part of the walk  $\gamma'$  has been removed, i.e., the identity

$$P(\gamma|\gamma') = P'(\gamma), \quad (\text{Markov property}) \quad (12)$$

where the prime refers to the new domain. The Markov property is self-evident for random walk and follows directly from Eq. (11), i.e.,  $P(\gamma|\gamma') = P(\gamma\gamma')/P(\gamma') = \mu^{-(L+L')}/\mu^{-L'} = \mu^{-L} = P'(\gamma)$ , where  $L$  and  $L'$  are the lengths of segments  $\gamma'$  and  $\gamma$ , respectively.

In the scaling limit the lattice distance goes to zero whereas the length of the walk diverges. Consequently, the distribution diverges and must be replaced by an appropriate probability measure [4,13,43]. However, in order to define the probability distribution or measure in the scaling limit we shall assume that the Markov property continues to hold and interpret the  $P$  in Eq. (12) as probability measures. The Markov property is essential in carrying over the lattice probability distributions in the scaling limit.

For the critical interfaces defined by an exploration processes in the case of site percolation, the Markov property follows by inspection since the propagation of the interface is entirely determined by the local process of occupying the next site with probability  $1/2$ . This locality property is specific for percolation which has a geometrical phase transition and does in the SLE context, to be discussed later, determine the scaling universality class.

In the case of an interface in the Ising model the Markov property also holds, but since the spins interact a little calculation is required [13]. Consider an interface  $\gamma$  defined by an exploration process. According to the rules of statistical mechanics the probability distribution for the interface  $\gamma$  is given by

$$P(\gamma) = \frac{Z(\gamma)}{Z}. \quad (13)$$

Here  $Z$  is the full partition function defined in Eq. (8) with appropriate boundary conditions imposed.  $Z(\gamma)$  is the partial partition function with the spins associated with the interface  $\gamma$  fixed,

$$Z(\gamma) = \sum_{\{\sigma_i\}, \gamma} \exp[-H(\gamma)/kT]. \quad (14)$$

The Hamiltonian  $H(\gamma)$  inferred from Eq. (6) is the energy of a spin configuration with the spins determining  $\gamma$  fixed;  $\{\sigma_i\}, \gamma$  indicate the configurations to be summed over. Evidently, we have the identity  $Z = \sum_{\gamma} Z(\gamma)$ , i. e.,  $\sum_{\gamma} P(\gamma) = 1$ .

Whereas in the random walk case we only considered an individual path and in the percolation case the interface only feels the nearby sites, an Ising interface is imbedded in the interacting spin systems and we have to define the Markov property more precisely with respect to a domain  $D$ . Consider an interface across the domain  $D$  from  $A$  to  $B$  and assume that the last part  $\gamma$  is conditioned on the determination of the first part  $\gamma'$ , i. e., given by the distribution  $P_D(\gamma|\gamma')$ . Next imagine that we cut the domain  $D$  along the interface  $\gamma'$ , i. e., break the interaction bonds between the spins determining  $\gamma'$ . The right and left face of  $\gamma'$  can then be considered part of the domain boundary and the Markov property states that the distribution of  $\gamma$  in the cut domain  $D \setminus \gamma'$  (i. e.,  $D$  minus  $\gamma'$ ) equals  $P_D(\gamma|\gamma')$ ,

$$P_D(\gamma|\gamma') = P_{D \setminus \gamma'}(\gamma). \quad (\text{Markov property}) \quad (15)$$

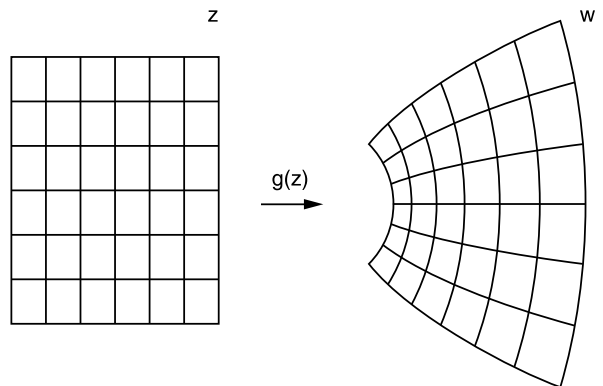
In order to demonstrate Eq. (15) we use the definition in Eq. (13). The conditional probability  $P_D(\gamma|\gamma') = P_D(\gamma\gamma')/P_D(\gamma') = (Z_D(\gamma\gamma')/Z_D)/(Z_D(\gamma')/Z_D) = Z_D(\gamma\gamma')/Z_D(\gamma')$ . Correspondingly, the distribution in the cut domain  $D \setminus \gamma'$  is  $P_{D \setminus \gamma'}(\gamma) = Z_{D \setminus \gamma'}(\gamma)/Z_{D \setminus \gamma'}$ . However, it follows from the structure of the partition function in Eqs. (6–8) that  $Z_{D \setminus \gamma'} = \exp[E(\gamma')/kT]Z_D(\gamma')$  and  $Z_{D \setminus \gamma'}(\gamma) = \exp[E(\gamma')/kT]Z_D(\gamma\gamma')$ , where  $E(\gamma')$  is the energy of the broken bonds. By insertion the interface Boltzmann factors cancel out and we obtain Eq. (15) expressing the Markov property.

### Conformal Invariance

In the complex plane analysis implies geometry. The representation of a complex number  $z = x + iy$  directly associates complex function theory with 2D geometrical shapes. This connection is of importance in mathematical physics in for example 2D electrostatics and 2D hydrodynamics. In the context of SLE Riemann's mapping theorem plays an essential role [1].

### Conformal Maps

Briefly, Riemann's mapping theorem [2] states that any simply connected domain, i. e., topologically deformable to a disk, in the complex plane  $z$  can be uniquely mapped to a unit disk  $|w| < 1$  in the complex  $w$  plane by mean of a complex function  $w = g(z)$ . By combining complex functions we can map any simply connected domain to any other simply connected domain. For example, if  $g_1(z)$  and  $g_2(z)$  map domains  $D_1$  and  $D_2$  to the unit disk, respectively, then  $g_2^{-1}(g_1(z))$  maps  $D_1$  to  $D_2$ ; here  $g_2^{-1}$  is the inverse function of  $g_2$ , i. e.,  $g_2^{-1}(g_2(z)) = z$ . As an example, the transformation  $g(z) = i(1+z)/(1-z)$  maps the unit disk centered at the origin to the upper half plane. Likewise, the Möbius transformation  $g(z) = (az+b)/(cz+d)$  determined by four real parameters,  $ad-bc > 0$ , maps the upper half plane onto itself. Conformal transformations are angle-preserving and basically correspond to a combination of a local rotation, local translation, and local dilatation. In terms of an elastic medium picture conformal transformations are equivalent to deformations without shear [54]. Expressing  $g(z)$  in terms of its real and imaginary parts,  $g(z) = u(x, y) + iv(x, y)$ , the Cauchy–Riemann equations  $\partial u/\partial x = \partial v/\partial y$  and  $\partial u/\partial y = -\partial v/\partial x$  hold implying that  $u$  and  $v$  are harmonic functions satisfying Laplace's equations  $\nabla^2 u = 0$  and  $\nabla^2 v = 0$ ,  $\nabla^2 = \partial^2/\partial x^2 + \partial^2/\partial y^2$ . In Fig. 3 we have depicted a conformal angle-preserving transformation effectuated by the complex function  $g(z)$  from the complex  $z$  plane to the complex  $w$  plane.



**Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 3**

We depict a conformal transformation from the complex  $z$  plane to the complex  $w$  plane. We note the angle-preserving property, i. e., a shear-free transformation. The map in the figure is given by  $w = z^2$

### Measures – Conformal Invariance

Whereas the Markov property discussed above holds for lattice curves even away from criticality, we here want to assume another property which only holds in the scaling limit at the critical point, namely conformal invariance. In the scaling limit we anticipate that the probability measure  $P(\gamma)$  for an interface  $\gamma$  is invariant under a conformal transformation. More precisely, consider a lattice model, say the Ising or percolation model, and specify two domains  $D$  and  $D'$  on the lattice. Next, consider an interface, cluster boundary or domain wall  $\gamma$  from the boundary points  $A$  and  $B$  across the domain  $D$ . In terms of the partition functions the probability distribution for  $\gamma$  is given by Eq. (13). We now perform the scaling or continuum limit of the lattice model keeping the domains  $D$  and  $D'$  fixed. The continuous random interface approaches its scaling form and is characterized by the measure  $P_D(\gamma)$ . At the critical point we assume that the interface is scale invariant under the larger symmetry of conformal transformations. The next step is to consider a specific conformal transformation  $g(z)$  which according to Riemann's theorem precisely maps domain  $D$  to domain  $D'$ , i. e.,  $D' = g(D)$  and the interface to  $g(\gamma)$ . The assumption of conformal invariance then states that the probability measure  $P$  is invariant under this transformation expressing the scale invariance, i. e.,

$$P_D(\gamma) = P_{g(D)}(g(\gamma)) . \quad (\text{Conformal property}) \quad (16)$$

Both the Markov property and the conformal property are sufficient in combination with Loewner evolution to determine the measures in the scaling limit.

### Loewner Evolution

The original motivation of Loewner's work was to examine the so called Bieberbach conjecture which states that  $|a_n| \leq n$  for the coefficients in the Taylor expansion  $f(z) = \sum_{n=0}^{\infty} a_n z^n$ , where  $f(z)$  maps the unit disk to the

complex plane. The conjecture was proposed in 1916 and finally proven in 1984 by de Branges [2,38,39]. For that purpose Loewner [64] considered growing parametrized conformal maps to a standard domain. In the present context Loewner's method allow us to access growing shapes in 2D in an indirect manner by means of a 'time dependent' conformal transformation  $g_t(z)$ .

### Growing Stick

Before we address the derivation of the Loewner equation let us consider the specific conformal transformation

$$g_t(z) = \sqrt{z^2 + 4t} . \quad (17)$$

For  $t = 0$  we have  $g_0(z) = z$ , i. e., the identity map. Likewise, for  $z \rightarrow \infty$  we obtain

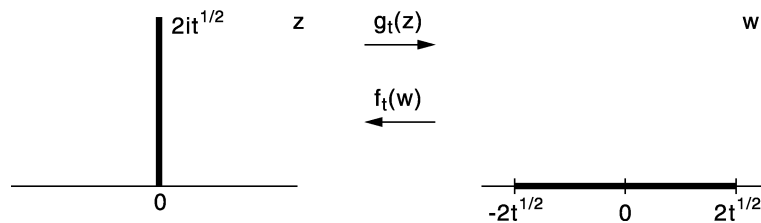
$$g_t(z) \sim z + \frac{2t}{z} , \quad (18)$$

showing that far away in the complex plane we again have the identity map. The coefficient in the next leading term,  $C_t/z$ , is called the capacity  $C_t$ ; here parametrized by the 'time variable',  $t = C_t/2$ . The map (17) has a branch point at  $z = 2it^{1/2}$  and it follows by inspection that  $g_t$  maps the upper half plane minus a stick from the origin  $0$  to  $2it^{1/2}$  back to the upper half plane. From the inverse map  $f_t(w) = g_t^{-1}(w)$ ,

$$f_t(w) = \sqrt{w^2 - 4t} , \quad (19)$$

we infer that the right face of the stick is mapped to the real axis from  $0$  to  $2t^{1/2}$ , the tip to the origin, and the left face to the interval  $-2t^{1/2}$  to  $0$ . Under the map the growing stick thus becomes part of the boundary in the  $w$  plane. The growing stick is depicted in Fig. 4.

The stick shows up as an imaginary contribution along the interval  $-2t^{1/2}$  to  $2t^{1/2}$ . More precisely, since  $f_t(w)$  is analytic in the upper half plane, implementing the asymptotic behavior  $f_t(w) \sim w$  for  $w \rightarrow \infty$ , and using Cauchy's



**Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 4**

We depict the growing stick corresponding to the conformal transformation  $w = \sqrt{z^2 + 4t}$ . The vertical cut in the  $z$  plane extends from the origin to the point  $(0, 2it^{1/2})$ . The right and left faces of the cut are mapped to the real axis from  $-2t^{1/2}$  to  $+2t^{1/2}$ , the endpoint to the origin, in the complex  $w$  plane

theorem, we obtain the dispersion relation or spectral representation

$$f_t(w) = w - \int \frac{d\omega}{\pi} \frac{A_t(\omega)}{w - \omega}, \quad (20)$$

with spectral weight  $A_t(\omega)$ . In the case of the growing stick we find  $A_t(\omega) = (4t - \omega^2)^{1/2}$  for  $\omega^2 < 4t$  and otherwise  $A_t(\omega) = 0$ . Using  $1/(\omega + i\epsilon) = \text{P} \, 1/\omega - i\pi\delta(\omega)$  (P denotes principal value) we also have  $\text{Im} f_t(\omega) = A_t(\omega)$  and  $\text{Re} f_t(\omega) = \omega - \text{P} \int (d\omega'/\pi) A_t(\omega')/(\omega - \omega')$ . The time dependent spectral weight  $A_t(\omega)$  thus characterizes the growing stick. With the chosen parametrization we also have the sum rule

$$\int \frac{d\omega}{\pi} A_t(\omega) = C_t = 2t. \quad (21)$$

Finally, we note that the map  $g_t$  satisfies the equation of motion

$$\frac{dg_t(z)}{dt} = \frac{2}{g_t(z)}, \quad (22)$$

i.e., solving Eq. (22) with the initial condition  $g_0(z) = z$  and the boundary condition  $g_t(z) \sim z$  for  $z \rightarrow \infty$  we arrive at the map in Eq. (17).

### Loewner Equation

The growing stick nicely illustrates the idea of accessing a growing shape indirectly by the application of Riemann's theorem mapping the domain adjacent to the shape to a standard reference domain, here the upper half plane. This so-called uniformizing map effectively absorbs the shape and encodes the information about the shape into the spectral weight  $A_t(\omega)$  along the real axis.

Let us consider a general shape or hull  $K_t$  in the upper half plane  $\mathbf{H}$ . Together with the real axis the shape forms part of the boundary of the domain  $\mathbf{D}$ . In other words, the domain in question is the upper half plane  $\mathbf{H}$  with the shape  $K_t$  subtracted,  $\mathbf{D} = \mathbf{H} \setminus K_t$ . Applying Riemann's theorem we map the simply connected domain  $\mathbf{D}$

back to the upper half plane  $\mathbf{H}$  by means of the conformal transformation  $g_t(z)$ , i.e.,  $g_t$  absorbs the shape  $K_t$ . Imagine that the shape grows a little bit further from  $K_t$  to  $K_{t+\delta t} = K_t + \delta K_t$ , where  $K_t$  is now part of  $K_{t+\delta t}$ ;  $\delta K_t$  is the shape increment. Correspondingly, the map  $g_{t+\delta t}$  is designed to absorb  $K_{t+\delta t}$ , i.e.,  $\mathbf{H} \setminus K_{t+\delta t} \rightarrow \mathbf{H}$  by means of the map  $g_{t+\delta t}$ . We now carry out the elimination in two ways. Either we absorb  $K_t$  by means of the map  $g_t$  and subsequently  $\delta K_t$  by means of the map  $\delta g_t$  or we absorb  $K_{t+\delta t}$  directly in one step by means of the map  $g_{t+\delta t}$ . Consequently, combining maps we have  $g_{t+\delta t}(z) = \delta g_t(g_t(z))$  or  $g_t(z) = \delta g_t^{-1}(g_{t+\delta t}(z))$ . Since  $\delta g_t^{-1}(w)$  is analytic in  $\mathbf{H}$  we obtain the spectral representation

$$\delta g_t^{-1}(w) = w - \int \frac{d\omega}{\pi} \frac{\delta A_t(\omega)}{w - \omega}, \quad (23)$$

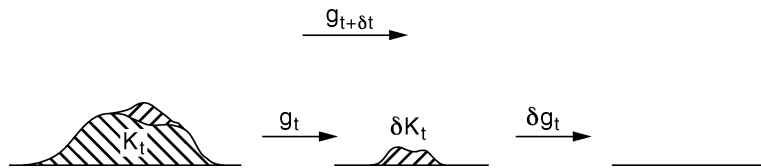
with infinitesimal spectral weight  $\delta A_t$ , or inserting  $w = g_{t+\delta t}(z)$

$$g_t(z) = g_{t+\delta t}(z) - \int \frac{d\omega}{\pi} \frac{\delta A_t(\omega)}{g_{t+\delta t}(z) - \omega}. \quad (24)$$

The last step is to set  $\delta A_t(\omega) = \rho_t(\omega)\delta t$ , yielding a differential equation for the evolution of the map  $g_t$  eliminating the shape  $K_t$ ,

$$\frac{dg_t(z)}{dt} = \int \frac{d\omega}{\pi} \frac{\rho_t(\omega)}{g_t(z) - \omega}. \quad (25)$$

Specifying the weight or measure  $\rho_t(\omega)$  along the real  $\omega$ -axis this equation determines, through the uniformizing map  $g_t$ , how the shape  $K_t$  grows. The spectral weight encodes the 2D shape into the real function  $\rho_t(\omega)$ . Note that since  $\rho_t(\omega)$  is not specified and can depend nonlinearly on the map, Eq. (25) still represents a highly nonlinear problem. Invoking the asymptotic condition  $g_t(z) \sim z + C_t/z$ , where  $C_t$  is the time dependent capacity we infer  $dC_t/dt = \int (d\omega/\pi) \rho_t(\omega)$  or since  $\rho_t(\omega) = dA_t(\omega)/dt$  the sum rule in Eq. (21). The conformal mapping procedure is depicted in Fig. 5.



**Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 5**

The combination of maps involved in the derivation of the Loewner equation. First the map  $g_t$  eliminates the hull  $K_t$ . Subject to the growth in the time interval  $\delta t$  the incremental hull  $\delta K_t$  is subsequently absorbed by the infinitesimal map  $\delta g_t$ . Correspondingly, the hull  $K_{t+\delta t}$  is absorbed by the map  $g_{t+\delta t}$  in one step

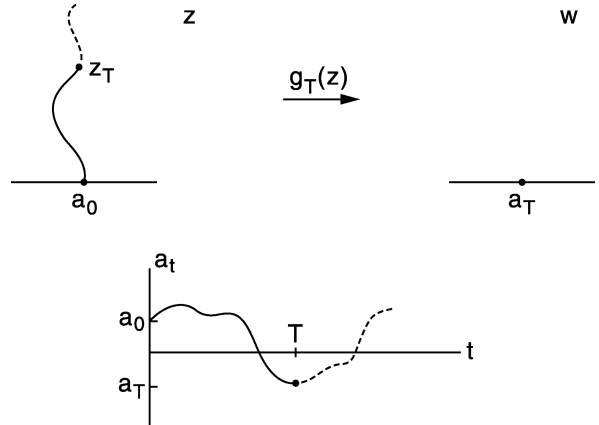
In the special case where the growth takes place at a point the equation simplifies considerably. Assuming that the spectral weight is concentrated at the point  $\omega = a_t$ , where  $a_t$  is a real function of  $t$  and setting  $\rho_t(\omega) = 2\pi\delta(\omega - a_t)$  we arrive at the Loewner equation

$$\frac{dg_t(z)}{dt} = \frac{2}{g_t - a_t}. \quad (26)$$

The Loewner equation describes the growth of a curve or trace  $\gamma_t$  with endpoint  $z_t$ ,  $0 < t < \infty$ , in the upper half complex  $z$ -plane. The time-dependent conformal transformation  $g_t$  maps the simply connected domain  $\mathbf{H} \setminus \gamma_t$ , i. e., the half plane excluding the curve  $\gamma_t$  back to the  $w$  half plane. At a given time instant  $t$  the tip of the curve  $z_t$  is determined by  $g_t(z_t) = a_t$ , i. e., the point where Eq. (26) develops a singularity. The topological properties and shape of the curve are encoded in the real function  $a_t$  which lives on the real axis in the  $w$ -plane. As  $a_t$  develops in time the tip of the curve  $z_t$  determined by  $z_t = g_t^{-1}(a_t)$  traces out a curve. Since the domain  $\mathbf{H} \setminus \gamma_t$  must be simply connected for the Riemann theorem to apply the curve or trace cannot cross itself or cross the real axis. Whenever the curve touches or intersect itself or the real axis the enclosed part will be excluded from the domain. In other, words, during the time progression the curve effectively absorbs part of the upper half plane. It is a deep property of Loewner evolution that the topological properties of a 2D non-crossing curve are entirely encoded by the real function  $a_t$ . The encoding works both ways: A given 2D non-crossing curve  $\gamma_t$  corresponds to a specific real function  $a_t$ , a given real function  $a_t$  yields a specific 2D non-crossing curve  $\gamma_t$ . A continuous  $a_t$  will yield a continuous curve  $\gamma_t$ . A discontinuous  $a_t$  in general gives rise to branching. Whether or not the curve intersects or touches itself is determined by the singularity structure of the drive  $a_t$ . In the case where the Hölder condition  $\lim_{\tau \rightarrow 0} |(a_{t+\tau} - a_t)/\tau^{1/2}|$  is greater than 4 we have self-intersection. Note again that since the curve is defined indirectly by the singularity structure in Eq. (26) we cannot easily identify a curve parametrization and for example determine a tangent vector, etc. The mechanism underlying the Loewner equation is shown in Fig. 6.

### Exact Solutions

In a series of simple cases one can solve the Loewner equation analytically [42,44]. For vanishing drive  $a_t = 0$  we obtain the growing vertical stick discussed above. Correspondingly, a constant drive  $a_t = a$  yields a vertical stick growing up from the point  $a$  on the real axis.



**Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 6**

The mechanism in the Loewner equation. The curve in the upper half complex  $z$  plane generated by the Loewner equation is mapped onto a finite but growing segment of the real axis of the complex  $w$  plane. The endpoint  $z_T$  is mapped to the real number  $a_T$ . As  $a_t$  develops in time and makes excursions along the real axis the endpoint  $z_t$  of the curve grows into the upper half plane

In the case of a linear drive  $a_t = t$  the tip of the curve  $z_t$  is given by  $z_t = 2 - 2\phi_t \cos \phi_t + 2i\phi_t$ , where the phase  $\phi_t$  is determined from the equations:  $2 \ln r_t - r_t \cos \phi_t = 2 \ln 2 + t - 2$  and  $r_t = 2\phi_t / \sin \phi_t$ . By inspection  $\phi_0 = 0$  and  $\phi_\infty = \pi$ . The curve thus approaches the asymptote  $2\pi i$  for  $t \rightarrow \infty$ . For small  $t$  analysis yields  $z_t \sim (2/3)t + 2i\sqrt{t}$ , i. e., the trace approaches the origin with infinite slope. The square root drives  $a_t = 2\sqrt{\kappa t}$  and  $a_t = 2\sqrt{\kappa(1-t)}$ ,  $0 < t < 1$  with a finite-time singularity can also be treated. In the first case,  $a_t = 2\sqrt{\kappa t}$ , the trace is a straight line  $z_t = B \exp(i\phi)\sqrt{t}$  forming the angle  $\phi$  with respect to the real axis. The amplitude  $B$  and phase  $\phi$  depend on the parameter  $\kappa$ . The angle  $\phi = (\pi/2)(1 - \kappa^{1/2}/(\kappa + 4)^{1/2})$ . For  $\kappa = 0$ ,  $\phi = \pi/2$  and we recover the perpendicular stick; for  $\kappa \rightarrow \infty$ ,  $\phi \rightarrow 0$  and the angle of intersection decreases to zero. In the second case,  $a_t = 2\sqrt{\kappa(1-t)}$ , the behavior of the trace is more complex. For  $0 < \kappa < 4$  the trace forms a finite spiral in the upper half plane; for  $\kappa = 4$  the trace has a glancing intersection with the real axis. For  $4 < \kappa < \infty$  the trace hits the real axis in accordance with the Hölder condition discussed above.

### Stochastic Loewner Evolution

After these preliminaries we are in position to address stochastic Loewner evolution (SLE). The essential observation made by Oded Schramm [77] within the context of



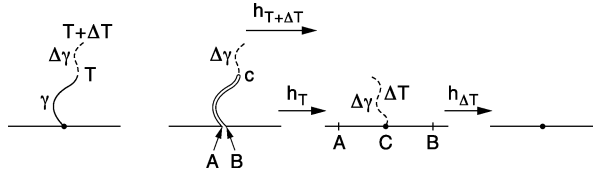
loop erased random walk was that the Markov and conformal properties of the measures or probability distributions for random curves generated by Loewner evolution imply that the random drive  $a_t$  must be proportional to an unbiased 1D Brownian motion.

### Schramm's Theorem

The Loewner Equation (26) generates a non-crossing curve in the upper half plane  $\mathbf{H}$  originating at the origin  $\mathbf{O}$ , given a continuous function  $a_t$  with initial value  $a_0 = 0$ . As  $a_t$  develops in time the tip of the curve  $z_t$  determined by the condition  $g_t(z_t) = a_t$  traces out a curve or trace. In the case where  $a_t$  is a continuous random function the Loewner Equation (26) likewise becomes a stochastic equation of motion yielding a stochastic map  $g_t(z)$ . As a result the trace determined by  $g_t(z_t) = a_t$  or  $z_t = g_t^{-1}(a_t)$  is a random curve. The issue is to establish a contact between the exploration processes defining interfaces in the lattice models, the scaling limit of these curves, and the curves generated by SLE. In the scaling limit we thus invoke the two properties discussed above: i) the Markov property in Eq. (12) and ii) the conformal property in Eq. (16).

In order to demonstrate the surprising property that the Markov and conformal properties in combination imply that  $a_t$  must be a 1D Brownian motion we focus on chordal SLE which applies to a random curve or trace connecting two boundary points. Since the probability distribution or measure  $P(\gamma)$  on the random curve  $\gamma$  using property ii) is assumed to be conformally invariant and since we by Riemann's theorem can map any simply connected domain to the upper half plane by mean of a conformal transformation, we are free to consider curves in the upper half plane from the origin  $\mathbf{O}$  to  $\infty$  parametrized with a time coordinate  $0 < t < \infty$ .

Imagine that we grow the curve from time  $t = 0$  to time  $T$  driven by the function  $a_t$ ,  $0 < t < T$ . The curve or trace is generated by the Loewner Equation (26) with boundary condition  $a_0 = 0$  and the trace  $z_t$  by  $g_t(z_t) = a_t$  or  $z_t = g_t^{-1}(a_t)$ . With the chosen time parametrization we have  $g_t(z) \sim z + 2t/z$  for  $z \rightarrow \infty$  in the upper half plane. The map  $g_t$  thus uniformizes the trace, i.e., the tip  $z_t$  is mapped to  $a_t$  on the real axis in the  $w$ -plane. In order to invoke the Markov property we let the curve grow the time increment  $\Delta T$  corresponding to the curve segment  $\Delta\gamma$ . The Markov property then implies that the distribution on  $\Delta\gamma$  conditioned on the distribution on  $\gamma$  is the same as the distribution on  $\Delta\gamma$  in the cut domain  $\mathbf{H} \setminus \gamma$ , i.e., the domain with the curve  $\gamma$  deleted; this stage is illustrated in Fig. 7.



**Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 7**

The figure depicts the construction in the derivation of SLE. The first step implements the Markov property by turning the curve  $\gamma$  into a cut. Subsequently, the conformal transformation  $h_T$  maps  $\gamma$  back to the origin. Finally, the map  $h_{\Delta T}$  maps the segment  $\Delta\gamma$  to the origin. The complete process is also implemented by  $h_{T+\Delta T}$ . The combination of the Markov property and conformal invariance implies that  $a_t$  performs a Brownian motion

Next, in order to implement the conformal property we shift the image by  $a_T$  in such a way that the curve segment  $\Delta\gamma$  again starts at the origin  $\mathbf{O}$ . This is achieved by using the map  $h_T = g_T - a_T$  which since  $g_T(z_T) = a_T$  maps the tip  $z_T$  back to the origin; this construction is also depicted in Fig. 7. Moreover, the asymptotic behavior of  $h_T$  for large  $z$  is  $h_T(z) \sim z - a_T - 2T/z$ . Since the measure by assumption is unchanged under the conformal transformation  $h_T$  we infer that  $\Delta\gamma$  growing the time  $\Delta T$  from the origin has the same distribution as the segment  $\Delta\gamma$  grown from time  $T$  to time  $T + \Delta T$  conditioned on the segment  $\gamma$  grown up to time  $T$ . Moreover, since the segment  $\gamma$  from  $\mathbf{O}(AB)$  to  $C$  subject to the Markov property has become part of the boundary, as shown in Fig. 7, we also infer that the measure on  $\Delta\gamma$  is independent of the measure on  $\gamma$ . Finally, applying  $h_{\Delta T}$  we map the segment  $\Delta\gamma$  to the origin as a common reference point as indicated in Fig. 7.

Since the random curves are determined by the random maps  $g_t$  and  $h_t$  driven by the random function  $a_t$  the issue is how to transfer the properties of the measure on the curve determined by the Markov and conformal properties to the measure on the random driving function  $a_t$ .

In order to combine the Markov properties arising from the analysis of the lattice models and the conformal invariance pertaining to critical random curves, we carry out the following steps. First we grow the curve  $\gamma$  from the origin  $\mathbf{O}$  to the tip  $z_T$ . Implying conformal invariance the curve is then uniformized back to the origin by means of  $h_T$ . The next step is to grow the curve segment  $\Delta\gamma$  in time  $\Delta T$ . This segment is subsequently absorbed by means of the map  $h_{\Delta T}$ . According to the Markov property the distribution of  $\Delta\gamma$  from  $\mathbf{O}$  to  $z_{\Delta T}$  is the same as the distribution of  $\Delta\gamma$  grown from  $T$  to  $T + \Delta T$  conditioned on  $\gamma$  grown from 0 to  $T$ . Since the curve  $\gamma$  is deter-

mined by the map  $h_t$  the distribution is reflected in  $h_t$ . In particular, the stochastic properties of the curve is transferred to the random function  $a_t$  generating the curve by the Loewner evolution. The last step is now to observe that absorbing the segment  $\Delta\gamma$  from  $\mathbf{O}$  to  $z_{\Delta T}$  by means of  $h_{\Delta T}$  is the same transformation as first applying the inverse map  $h_T^{-1}$  followed by the map  $h_{T+\Delta T}$ ; in both cases the end result is the absorption of the initial curve  $\gamma + \Delta\gamma$ , see Fig. 7. As regards the measure or distribution we have the equivalence  $h_{\Delta T}(z) \approx h_{T+\Delta T}(h_T^{-1}(z))$ . Using the asymptotic form  $h_t(z) \sim z - a_t - 2t/z$  we obtain  $a_{T+\Delta T} - a_T \approx a_{\Delta T}$ ; note that  $\approx$  indicates identical distributions or measures.

In conclusion, the Markov property in combination with conformal invariance implies that  $a_{T+\Delta T} - a_T$  is distributed like  $a_{\Delta T}$  (stationarity) and that  $a_{\Delta T}$  and  $a_{\Delta T'}$  are independently distributed for non-overlapping time intervals  $\Delta T$  and  $\Delta T'$  (Markov property). Referring to Sect. “Scaling”, A on Brownian motion as expressed in Eqs. (3) and (4), i.e., stationarity and independence, we infer that  $a_t$  is proportional to a Brownian motion of arbitrary strength  $\kappa$ , i.e.,  $a_t = \sqrt{\kappa}B_t$ . Note that the reflection symmetry  $x \rightarrow -x$  holding in the present context rules out a bias or drift in the 1D Brownian motion [13,27].

This is the basic conclusion reached by Schramm in the context of loop erased random walk. Driving Loewner evolution by means of 1D Brownian motion with different diffusion coefficient or strength  $\kappa$  we generate a one-parameter family of conformally invariant or scale invariant non-crossing random curves in the plane.

### SLE Properties

Stochastic Loewner evolution is determined by the nonlinear stochastic equation of motion

$$\frac{dg_t}{dt} = \frac{2}{g_t - a_t}, \quad a_t = \sqrt{\kappa}B_t. \quad (27)$$

In the course of time  $a_t$  performs a 1D Brownian motion on the real axis starting at the origin  $a_0 = 0$ .  $a_t$  is a random continuous function of  $t$  and distributed according to  $\sqrt{\kappa}B_t$ . More precisely,  $a_t$  is given by the Gaussian distribution

$$P(a, t) = (2\pi t)^{-1/2} \exp[-a^2/2\kappa t], \quad (28)$$

with correlations

$$\langle (a_t - a_s)^2 \rangle = \kappa |t - s|. \quad (29)$$

First we notice that a constant shift of the drive  $a_t$ ,  $a_t \rightarrow a_t + b$ , is readily absorbed by a corresponding

shift of the map,  $g_t \rightarrow g_t + b$ . Moreover, using the scaling property of Brownian motion,  $B_{\lambda^2 t} = \lambda B_t$ , following from e.g. Eq. (5), we have  $a_{\lambda^2 t} = \lambda a_t$  and we conclude from Eq. (27) that  $g_t(z)$  has the same distribution as  $(1/\lambda)g_{\lambda^2 t}(\lambda z)$ , i.e.,  $g_{\lambda^2 t}(\gamma z) \approx \gamma g_t(z)$ . Note that this dilation invariance is consistent since the origin  $z = 0$  and  $z = \infty$ , the endpoints of curves, are preserved. Note also that the strength of the drive  $\kappa$  is an essential parameter which cannot be scaled away.

### Curves – Hulls – Bessel Process

For vanishing drive,  $\kappa = 0$ , the SLE yields a non random vertical line from  $z = 0$ , i.e., the growing stick discussed in Subsect. “Growing Stick”. As we increase  $\kappa$  the curve becomes random with excursions to the right and to the left in the upper half plane. Up to a critical value of  $\kappa$  the random curve is simple; i.e., non-touching or non self-intersecting. At a critical value of  $\kappa$  the Brownian drive is so strong that the curve begins to intersect itself and the real axis. These intersection take place on all scales since the curve is self-similar or scale invariant. Denoting the curve by  $\gamma_t$  we observe that since Riemann’s theorem uniformizing  $\mathbf{H} \setminus \gamma_t$  to  $\mathbf{H}$  only applies to a simply connected domain, the regions enclosed by the self-intersections do not become uniformized but are effectively removed from  $\mathbf{H}$ . The curve  $\gamma_t$  together with the enclosed parts is called the hull  $\mathbf{K}_t$  and the mapping theorem applies to  $\mathbf{H} \setminus \mathbf{K}_t$ .

In order to analyze the critical value of  $\kappa$  we consider the stochastic equation for  $h_t(z) = g_t(z) - a_t$ . From Eq. (27) it follows that

$$\frac{dh_t}{dt} = \frac{2}{h_t} + \xi_t, \quad (30)$$

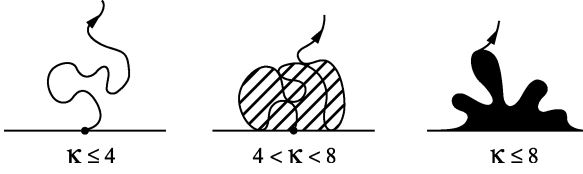
where  $\xi_t = -da_t/dt$  is white noise with correlations  $\langle \xi_t \xi_s \rangle = \kappa \delta(t - s)$ .

The nonlinear complex Langevin Equation (30) maps the tip of the curve  $z_t$  back to the origin. Likewise  $\mathbf{H} \setminus \gamma_t$  is mapped onto  $\mathbf{H}$ . A point  $x$  on the real axis is mapped to  $x_t = h_t(x)$  where  $x_t$  satisfies

$$\frac{dx_t}{dt} = \frac{2}{x_t} + \xi_t. \quad (31)$$

The Langevin Equation (31) is known as the Bessel equation and governs the radial distance  $R$  from the origin of a Brownian particle in  $d$  dimensions.

Introducing  $R = (\sum_{i=1}^d B_i^2)^{1/2}$  where  $B_i$ ,  $i = 1, \dots, d$ , is a 1D Brownian motion with distribution  $P(B, t) = (2\pi\kappa t)^{-1/2} \exp[-B^2/2\kappa t]$ , we find  $P(R, t) \propto (2\pi\kappa t)^{-d}$



**Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 8**

The figures depict the phases of SLE. For  $\kappa \leq 4$  the SLE trace is a simple non-intersecting scale invariant random curve from the origin to infinity with a fractal dimension between 1 and  $3/2$ . For  $4 < \kappa \leq 8$  the SLE curve is self-intersecting on all scales and also intersects the real axis on all scales. The curve together with the enclosed regions, the hull, eventually exhausts the upper half plane. The scale invariant hull has a fractal dimension ranging between  $3/2$  and 2. For  $\kappa \geq 8$  the fractal dimension of the hull locks onto 2 and the scale invariant hull is dense and plane-filling

$R^{d-1} \exp[-R^2/2\kappa t]$  satisfying the Fokker-Planck equation  $\partial P/\partial t = (\kappa/2)\partial^2 P/\partial R^2 - (\kappa(d-1)/2)\partial(P/R)/\partial R$ , corresponding to the Langevin equation

$$\frac{dR}{dt} = \kappa \frac{d-1}{2R} + \xi_t. \quad (32)$$

For  $d \leq 2$  Brownian motion is recurrent, i.e., the particle returns to the origin  $R = 0$ , for  $d > 2$  the particle goes off to infinity. Referring to Eq. (31) and setting  $\kappa(d-1)/2 = 2$  we obtain  $\kappa = 4/(d-1)$ , i.e.,  $R \rightarrow \infty$  for  $\kappa < 4$  and  $R \rightarrow 0$  for  $\kappa > 4$ .

Since the tip of the curve  $z_t$  is mapped to  $a_t$ , i.e.,  $h(z_t) = 0$ , the case  $x_t \rightarrow \infty$  for  $\kappa < 4$  corresponds to a curve never intersecting the real axis, i.e., the curve is simple. For  $\kappa > 4$  we have  $x_t \rightarrow 0$  corresponding to the case where the tip  $z_t$  intersects the real axis forming a hull. Since the curve is self-similar the intersections takes place on all scales and eventually the whole upper half plane is engulfed by the hull.

The marginal value  $\kappa = 4$  can also be inferred from a simple heuristic argument [27]. For small  $\kappa$  we can ignore the noise in Eq. (31) and the particle is repelled according to the solution  $x_t^2 \sim 4t$ . For large noise we ignore the nonlinear term and the noise can drive  $x_t$  to zero; we have  $x_t^2 \sim \kappa t$ . The balance is obtained for  $\kappa = 4$ .

In conclusion, for  $\kappa \leq 4$  the curve is simple, for  $\kappa > 4$  the curve intersects itself and the real axis infinitely many times on all scales, eventually the hull swallows the whole plane. For large  $\kappa$  the trace turns out to be plane-filling. The two cases  $\kappa \leq 4$  and  $\kappa > 4$  are depicted in Fig. 8.

## Fractal Dimension

The SLE random curves are fractal. An important issue is thus the determination of the fractal dimension  $D$  in terms of the Brownian strength or SLE parameter  $\kappa$ . It has been shown that [15,16,74]

$$D = 1 + \frac{\kappa}{8} \quad \text{for } 0 \leq \kappa \leq 8, \quad (33)$$

for  $\kappa \geq 8$  the fractal dimension locks onto 2 and the SLE curve is plane-filling.

In order to illustrate a typical SLE calculation we follow Cardy [27] in a heuristic derivation of Eq. (33). In order to evaluate  $D$  and in accordance with its definition [32,68] the standard procedure is to cover the object with disks of size  $\epsilon$  and follow how the number of disks  $N(\epsilon)$  of size  $\epsilon$  scales with  $\epsilon$  for small  $\epsilon$ , i.e.,  $N(\epsilon) \propto \epsilon^{-D}$ . However, since the SLE curve is random the argument has to be rephrased. Alternatively, we consider a disc of size  $\epsilon$  located at a fixed position  $z$  and ask for the probability  $P(z, \epsilon)$  that the SLE curve crosses the disk. The number of disks covering an area  $A$  is  $N_A = A/\epsilon^2$ , i.e.,  $P \propto \epsilon^{-D}/N_A$ , and we infer  $P(z, \epsilon) \propto \epsilon^{2-D}$ , where  $D$  is the fractal dimension. Incorporating the Markov property we subject the curve to an infinitesimal conformal transformation,  $h_{\delta t} = g_{\delta t} - a_{\delta t}$ , transforming the point  $z$  to  $w = g_{\delta t}(z) - a_{\delta t}$ ; moreover, all lengths are scaled by  $|h'_{\delta t}(z)|$  [1]. Setting  $z = x + iy$  and  $w = x' + iy'$  we obtain expanding Eq. (30)  $x' + iy' = 2\delta t/(x + iy) - \sqrt{\kappa}\delta B_t$  or  $x' = x + 2x\delta t/(x^2 + y^2) - \sqrt{\kappa}\delta B_t$ ,  $y' = y - 2y\delta t/(x^2 + y^2)$ ,  $\epsilon' = (1 - |h_{\delta t}(z)|)\epsilon$ , and  $|h_{\delta t}| = 2(x^2 - y^2)/(x^2 + y^2)^2$ . By conformal invariance the probability measure  $P(x, y, \epsilon)$  is unchanged and we infer

$$P(x, y, \epsilon) = \langle P(x', y', \epsilon') \rangle_{\delta B}, \quad (34)$$

where we have averaged over the Brownian motion referring to the initial part of the curve which has been eliminated by the map  $h_{\delta t}$ . Expanding Eq. (34) to first order in  $\delta t$  and noting that  $\langle (\delta B_t)^2 \rangle = \delta t$  we arrive at a partial differential equation for  $P(x, y, \epsilon)$

$$\left( \frac{2x}{x^2 + y^2} \frac{\partial}{\partial x} - \frac{2y}{x^2 + y^2} \frac{\partial}{\partial y} + \frac{\kappa}{2} \frac{\partial^2}{\partial x^2} - \frac{2(x^2 - y^2)}{(x^2 + y^2)^2} \epsilon \frac{\partial}{\partial \epsilon} \right) P = 0. \quad (35)$$

Since  $P(x, y, \epsilon) \propto \epsilon^{2-D}$  we have  $\epsilon \partial P/\partial \epsilon = (2-D)P$  and the determination of  $D$  is reduced to an eigenvalue problem. By inspection one finds

$$P \propto \epsilon^{1-\kappa/8} y^{(\kappa-8)^2/8\kappa} (x^2 + y^2)^{(\kappa-8)/2\kappa}, \quad (36)$$

and we identify the fractal dimension  $D = 1 + \kappa/8$  for  $\kappa < 8$ ; for  $\kappa > 8$  another solution yields  $D = 2$ .

## Results and Discussion

Stochastic Loewner evolution based on Eq. (27) generates conformally invariant non-crossing random curves in the upper half plane starting at the origin and going off to infinity. This is the case of chordal SLE, where the random curve connects two boundary points (the origin  $\mathbf{O}$  and infinity  $\infty$ ). Another case is radial SLE for random curves connecting a boundary point and an interior point in a simply connected domain [13,27,43]. Radial SLE is governed by another stochastic equation and will not be discussed here.

### Phase Transitions – Locality – Restriction – Duality

**Phase Transitions** SLE exhibits two phase transitions; for  $\kappa = 4$  and  $\kappa = 8$ . For  $0 < \kappa \leq 4$  the random curve is non-intersecting, i. e., a simple random continuous curve from  $\mathbf{O}$  to  $\infty$ . For  $4 < \kappa \leq 8$  the curve is self-intersecting on all scales. The curve together with the excluded regions form a hull which in the course of time absorbs the upper half plane. For  $\kappa$  just above 4 the half plane is eventually absorbed but the trace does not visit all regions, i. e., the hull is not dense. As we approach  $\kappa = 8$  the trace becomes more dense and the hull becomes plane-filling. This is also reflected in the fractal dimension  $D = 1 + \kappa/8$ . For  $\kappa > 8$  the hull is plane-filling, i. e.,  $D = 2$ . As we increase the strength of the Brownian drive further the excursions of the trace to the right and left in the upper half plane become more pronounced and the hull becomes vertically compressed. These results have been obtained by Rohde and Schramm [74] and Lawler et al. [57]. The various phases of SLE are depicted in Fig. 8. In Fig. 9 we have depicted numerical renderings of SLE traces for various values of  $\kappa$  (with permission from V. Beffara, <http://www.umpa.ens-lyon.fr/~vbeffara/simu.php>).

**Locality – Restriction** In addition to the phase transitions at  $\kappa = 4$  and  $\kappa = 8$ , there are special values of  $\kappa$  where SLE shows a behavior characteristic of the scaling limit of specific lattice models: The locality property for  $\kappa = 6$  and the restriction property for  $\kappa = 8/3$ . The issue here is the influence of the boundary on the SLE trace.

To illustrate the locality property, consider for example the SLE trace originating at the origin and purporting to describe the scaling limit of a domain wall in the lattice model. Due to the long range correlations at the critical point it is intuitively clear that a deformation of the boundary, e. g., a bulge  $\mathbf{L}$  on the real axis to the right of the origin, will influence the trace and push it to the left. A detailed analysis show that only for  $\kappa = 6$  is the trace

independent of a change of the boundary, i. e., the trace does not feel the boundary until it encounters a boundary point [59,63]. Returning to the lattice models the locality property for  $\kappa = 6$  applies specifically to the percolation case where the interface generated by the exploration process is governed by a local rule and the model has a geometric phase transition.

The restriction property is less obvious to visualize but basically states that the distribution of traces conditioned not to hit a bulge  $\mathbf{L}$  on the real axis away from the origin is the same as the distribution of traces in the domain where  $\mathbf{L}$  is part of the boundary, i. e., in the domain  $\mathbf{H} \setminus \mathbf{L}$ . Analysis shows that the restriction property only applies in the case for  $\kappa = 8/3$ . Among the lattice models only the scaling limit of self-avoiding random walk (SAW), where the measure is uniform, conforms to the restriction property and thus corresponds to  $\kappa = 8/3$  [63].

**Duality** For  $\kappa > 4$  the SLE generates a hull of fractal dimension  $D > 3/2$ . The boundary, external perimeter, or frontier of the hull is again a simple conformally invariant random curve characterized by the fractal dimension  $\bar{D}$ . Using methods from 2D quantum gravity Duplantier [31] has proposed the relationship,

$$(D - 1)(\bar{D} - 1) = \frac{1}{4}, \quad (37)$$

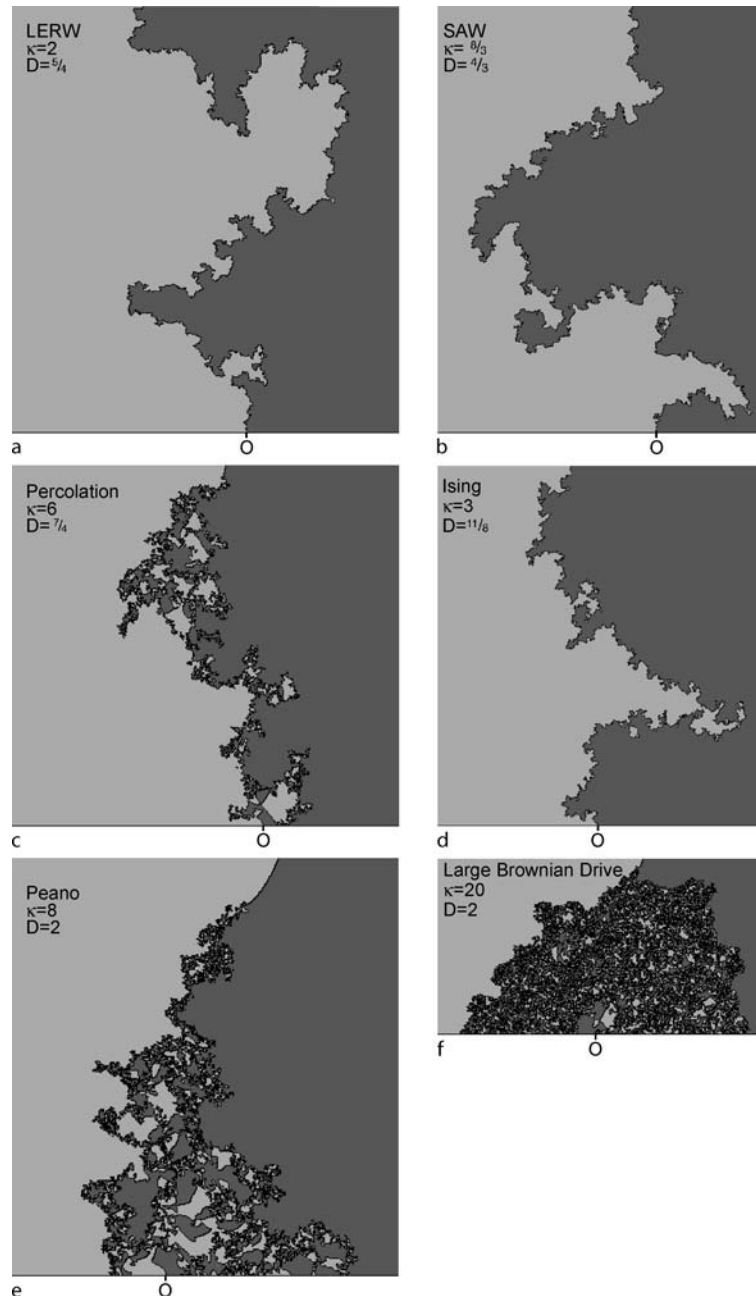
between the fractal dimension of the hull and its frontier. This result has been proved by Beffara for  $\kappa = 6$ , i. e., the percolation case [16]. Inserting in Eq. (33) we obtain for the corresponding SLE parameter the duality relation

$$\kappa \bar{\kappa} = 16. \quad (38)$$

### Loop Erased Random Walk (LERW)

Whereas the scaling limit of random walk, i. e., Brownian motion, does not fall in the SLE category because of self-crossings rendering Riemann's mapping theorem inapplicable, variations of Brownian motion are described by SLE.

Loop erased random walk (LERW) where loops are removed along the way is by construction self-avoiding and was introduced as a simple model of a self-avoiding random walk. LERW was studied by Schramm in his pioneering work [77]. LERW has the Markov property and has been proved to be conformally invariant in the scaling limit and described by SLE for  $\kappa = 2$  [57]. According to Eq. (33) LERW has the fractal dimension  $D = 5/4$  [67]. Also, since  $\kappa < 4$  LERW is non-intersecting. A simulation of LERW based on SLE is shown in Fig. 9a.



Stochastic Loewner Evolution: Linking Universality, Criticality and Conformal Invariance in Complex Systems, Figure 9

We depict a numerical rendering of SLE for a variety of  $\kappa$  values. In a we show loop erased random walk (LERW) for  $\kappa = 2$  with fractal dimension  $D = 5/4$ . In b we illustrate the case of self-avoiding random walk (SAW) for  $\kappa = 8/3$  and fractal dimension  $D = 4/3$ ; both LERW and SAW have  $\kappa > 4$  and are simple scale invariant random curves. In c we depict site percolation for  $\kappa = 6$  with fractal dimension  $D = 7/4$ . Since  $\kappa > 4$  the percolation case is self-intersecting and duality implies that the boundary or frontier of the hull is described by a SLE curve for  $\kappa = 16/6 = 8/3$ , i. e., the case of SAW. In d we show the Ising case for  $\kappa = 3$  and fractal dimension  $D = 11/8$ . In e we depict the limiting case  $\kappa = 8$  and fractal dimension  $D = 2$ . The hull is dense and plane-filling. The frontier of the hull corresponds to the SLE case  $\kappa = 16/8 = 2$ , i. e., the case of LERW. The so-called uniform spanning tree (UST) has the same properties as LERW and the SLE case for  $\kappa = 8$  can thus be thought of as a random plane filling Peano curve wrapping around the UST. Finally, in f we show the SLE trace and hull for  $\kappa = 20$  and  $D = 2$ . Because of the large Brownian excursions the plane-filling hull is vertically compressed (with permission from V. Beffara: <http://www.umpa.ens-lyon.fr/~vbeffara/simu.php>)



### Self-Avoiding Random Walk (SAW)

Self-avoiding random walk (SAW) is a random walk conditioned not to cross itself. SAW has been used to model polymers in a dilute solution and has a uniform probability measure. Since SAW satisfies the restriction property it is conjectured in the scaling limit to fall in the SLE class with  $\kappa = 8/3$  [46,47,48,61], yielding the fractal dimension  $D = 4/3$ . We note that Flory's mean field theory [30] for the size  $R$  of a polymer composed of  $N$  links (monomers) scales like  $R \sim N^\nu$ , where  $\nu = 3/(2+d)$  for  $d \leq 4$ . By a box covering we infer  $N \sim R^D$  where  $D$  is the fractal dimension, i. e.,  $D = (2+d)/3$ . In  $d = 2$  we obtain  $D = 4/3$  in accordance with the SLE result. SLE induced SAW in the scaling limit is shown in Fig. 9b.

### Percolation

The scaling limit of site percolation was conjectured by Schramm [77] to fall in the SLE class for  $\kappa = 6$ . Subsequently, the scaling limit of site percolation on a triangular lattice has been proved by Smirnov [78,80]. Percolation exhibits a geometrical phase transition. In the exploration process defining a critical interface the rule for propagation is entirely local. The lack of stiffness as for example in the Ising case to be discussed below results in a strongly meandering path winding back and in the scaling limit intersecting earlier part of the path. Since  $\kappa > 4$  the path together with the enclosed part, i. e., the hull, eliminates the whole plane in the course of time. As discussed above the locality property is specific to percolation and yields  $\kappa = 6$ . The fractal dimension of the percolation interface is according to Eq. (33)  $D = 7/4$ . We note that  $D$  is close to 2, i. e., the percolation interface nearly covers the plane densely. A series of new results and a proof of Cardy's conjectured formula for the crossing probability have appeared; we refer to [13,27,43] for details. Using the duality relation (38) the frontier of the percolation hull is a simple SLE curve for  $\kappa = 8/3$ , corresponding to SAW. In Fig. 9c we have depicted a SLE generated percolation interface.

### Ising Model – $O(n)$ Models

The Ising model in Eq. (6) is a special case of the  $O(n)$  model defined by the Hamiltonian

$$H = -J \sum_{\langle ij \rangle} \vec{\sigma}_i \vec{\sigma}_j, \quad (39)$$

where  $\vec{\sigma}_i = (\sigma_1, \dots, \sigma_n)$  is an  $n$ -component unit vector associated with the site  $i$ . For  $n = 1$  we recover the Ising model,  $n = 2$  is the XY-model [29], and  $n = 3$  the Heisenberg model.

By means of the Fortuin–Kasteleyn (FK) transformation based on a high temperature expansion the configurations of the  $O(n)$  model can be described by clusters or graphs on a dual lattice [35]. The crossing domain wall in Fig. 1 is thus a special case of a FK graph if we interpret the representation as a triangular Ising model. It has been conjectured that  $n$  is related to the SLE parameter  $\kappa$  by

$$n = -2 \cos(4\pi/\kappa) \quad \text{for } 8/3 \leq \kappa \leq 4. \quad (40)$$

In the Ising case  $n = 1$  and we have  $\kappa = 3$  yielding the fractal dimension  $D = 11/8$  for the Ising domain wall [84]. Since  $\kappa < 4$  the Ising domain wall is non-intersecting. Unlike the percolation case, the Ising interface is stiffer due to the interaction. We also note that the scaling limit of spin cluster boundaries in the Ising model recently has been proved to correspond to SLE for  $\kappa = 3$  [79]. The interface is shown in Fig. 9d.

### SLE – Conformal Field Theory

Whereas conformal field theory (CFT) is based on the concept of a local field  $\phi(r)$  and its correlations and therefore only accesses the underlying geometry indirectly through field correlations, SLE directly produces conformally invariant geometrical objects. A major issue is therefore the connection between CFT and SLE [6,7,26]. In CFT the central charge  $c$  plays an important role in delimiting the universality classes of the variety of lattice models yielding conformal field theories in the scaling limit. Percolation thus corresponds to the central charge  $c = 0$ , whereas the Ising model is associated with the central charge  $c = 1/2$ . It has been conjectured that the connection between the SLE parameter  $\kappa$  and the central charge  $c$  is given by

$$c = \frac{(6-\kappa)(3\kappa-8)}{2\kappa} = 1 - 6 \frac{(\kappa-4)^2}{4\kappa}. \quad (41)$$

We note that  $c < 1$  and, moreover, invariant under the duality transformation  $\kappa \rightarrow 16/\kappa$ .

### SLE – 2D Turbulence

There is an interesting application of SLE ideas in the context of 2D turbulence. The issue here is to analyze conformal invariance by comparing the statistical properties of geometrical shapes like domain walls with SLE traces with the view of determining the SLE parameter  $\kappa$  and the corresponding universality class.

In 3D turbulence is governed by the incompressible Navies–Stokes equation for the velocity field. Since the viscosity is only effective at small length scales 3D turbulence is characterized by a cascade of kinetic energy  $(1/2)v^2$  from large scales (driving scale) to small scales (dissipation scale). In the inertial regime the energy spectrum  $E(k)$  ( $k$  is the wavenumber) is characterized by the celebrated Kolmogorov 5/3 law [46],  $E(k) \propto k^{-5/3}$ , indicating an underlying scale invariance in turbulence.

In 2D the cascade picture is different. Since both kinetic energy and squared vorticity (enstrophy) are conserved in the absence of dissipation and forcing, two cascades coexist [52,53]. A direct cascade to small scales for the squared vorticity  $\omega^2 = (\nabla \times v)^2$  with scaling exponent  $-3$  and an inverse cascade to larger scales for the kinetic energy  $(1/2)v^2$  with Kolmogoroff scaling exponent  $-5/3$ . The system is thus characterized by a fine scale vorticity structure together with a large scale velocity structure. Moreover, we can assume that the vorticity structure is equipartitioned, i.e., in equilibrium.

In order to investigate whether the scale invariance of the small scale vorticity structure can be extended to conformal invariance Bernard et al. [17] have considered the statistics of the boundaries of vorticity clusters. By comparing the zero-vorticity isolines with SLE traces they find that cluster boundaries fall in the universality class corresponding to  $\kappa = 6$ , i.e., the case of percolation. Since 2D turbulence is a driven nonequilibrium system, this observation is very intriguing in particular since the correlations between vortices are long-ranged. A similar analysis [19] of the isolines in the inverse cascade in surface quasi-geostrophic turbulence corresponds to  $\kappa = 4$ , i.e., from Eq. (40) the domain walls in the equilibrium XY model for  $n = 2$ . For comments on the application of SLE in turbulence we refer to Cardy [28].

### SLE – 2D Spin Glass

It is a standing issue whether conformal field theory can be applied to disordered systems, in particular systems with quenched disorder. In recent work Amoruso et al. [3] and Bernard et al. [18] have considered zero temperature domain walls in the Ising spin glass [34]; see also [33]. The Ising spin glass is an equilibrium system with quenched disorder. The system is described by the Hamiltonian  $H = -\sum_{\langle ij \rangle} J_{ij} \sigma_i \sigma_j$ , where the random exchange constants  $J_{ij}$  are picked from a Gaussian distribution with zero mean. The glass transition is at  $T = 0$  and the system has a two-fold degenerate ground state. Inducing a scale invariant domain wall between the two ground states and comparing with an SLE trace, it is found that both the

Markov and conformal properties are obeyed and that the universality class corresponds to  $\kappa \approx 2.3$ .

### Further Remarks

In this discussion we have left out several topics which have played an important role in the development and applications of SLE. We mention some of them below.

There is an interesting connection between LERW and the so-called uniform spanning tree (UST) [13,43,57,77]. A spanning tree is a collection of vertices and edges which form a tree, i.e., without loops or cycles. A uniform spanning tree is a random spanning tree picked among all possible spanning trees with equal probability. Consider the unique path between two vertices on a UST. Since the path lives on a tree it is by construction non-crossing and it turns out that it has the same distribution as LERW. The winding random curve enclosing the UST can be visualized as a random plane-filling Peano curve. In the scaling limit the Peano curve is described by SLE for  $\kappa = 8$  with fractal dimension  $D = 2$ .

The  $q$ -state Potts model [87] constitutes a generalization of the Ising model; here the lattice variable takes  $q$  values. The model is defined by the Hamiltonian  $H = -J \sum_{\langle ij \rangle} \delta_{\sigma_i \sigma_j}$ , where  $\sigma_i = 1, \dots, q$ ; the Ising model obtains for  $q = 2$ . Applying the high temperature FK representation the configurations can be represented by loops and domain walls. From considerations involving the fractal dimension [76] it has been conjectured that domain walls in the scaling limit of the Potts model fall in the SLE category for  $q = 2 + 2 \cos(8\pi/\kappa)$ , where  $4 \leq \kappa \leq 8$ . For  $q = 2$  we recover the Ising case for  $\kappa = 3$ . In the limit  $q \rightarrow 0$ , the graph representation is equivalent to the uniform spanning tree described by SLE for  $\kappa = 8$ . For a numerical study of the three-state Potts model and its relation to SLE consult [36].

Standard SLE is driven by 1D Brownian motion producing a fractal curve. Ruskin et al. [75] have considered the case of adding a stable Lévy process with shape parameter  $\alpha$  to the Brownian motion. Backing their analysis with numerics they find that the SLE trace branches and exhibit a ‘phase transitions’ related to self-intersections.

2D Brownian motion, the scaling limit of 2D random walk, is an incredibly complex fractal coil owing to the self-crossings on all scales. Although 2D Brownian motion because of self-crossing itself falls outside the SLE scheme, the outer frontier or perimeter of 2D random walk is a non-crossing and non-intersecting fractal curve which can be accessed by SLE. Verifying an earlier conjecture by Mandelbrot [68] it has been proven using SLE techniques [58] that the fractal dimension of the Brownian

perimeter is  $D = 4/3$ , i.e., the same as the fractal dimension of self-avoiding random walk and the external perimeter of the percolation hull. Other characteristics of Brownian motion such as intersection exponents have also been obtained [59,60,62]; see also [66].

In recent work Zoia et al. [88] have considered the distribution of first passage times and distances along critical curves generated by SLE for different values of  $\kappa$ .

### Future Directions

Stochastic Loewner evolution represents a major step in our understanding of fractal shapes in the 2D continuum limit. By combining the Markov property (stationarity) with conformal invariance SLE provides a minimal scheme for the generation of a one-parameter family of fractal curves. The SLE scheme also provides calculational tools which have led to a host of new results. SLE is a developing field and we can on the mathematical front anticipate progress and proofs of some yet unproved scaling limits, e. g., the scaling limit of the FK representation of the Potts model and the scaling limit of SAW.

On the more physical front many issues also remain open. First there is the fundamental issue of the connection between the hugely successful but non-rigorous CFT and SLE. Here progress is already under way. In a series of papers Bauer and Bernard [6,7,8,9,10,12] have shown how SLE results can be derived using CFT methods. Cardy [26] have considered a multiple SLE process and the connection to Dyson's Brownian process and random matrix theory. The analysis of the CFT-SLE connection still remains to be investigated further.

An obvious limitation of SLE is that it only addresses critical domain walls and not the full configuration of clusters and loops in for example the FK representation of the Potts model. In the case of critical percolation this problem has been addressed by Camia and Newman [21]. Another issue is how to provide SLE insight into spin correlations in the Potts or  $O(n)$  models.

In the original formulation of SLE the Markov and conformal properties essentially requires a Brownian drive. It is clearly of interest to investigate the properties of random curves generated by other random drives. Such a program has been initiated by Ruskin et al. [75] who considered adding a Lévy drive to the Brownian drive; see also work by Kennedy [49,50].

Since the SLE trace lives in the infinite upper half plane the whole issue of finite size effects remain open. In ordinary critical phenomena the concept of a Kadanoff block construction and the diverging correlation length near the transition lead to a theory of finite size scaling and cor-

rections to scaling which can be accessed numerically. It is an open problem how to develop a similar scheme for SLE.

In statistical physics it is customary and natural to associate a free energy to a domain wall and an interaction energy associated with several domain walls. These free energy considerations are entirely absent in the SLE framework which is based on conformal transformations. A major issue is thus: Where is the free energy in all this and how do we reintroduce and make use of ordinary physical considerations and estimates [69].

### Acknowledgments

The present work has been supported by the Danish Natural Science Research Council.

### Bibliography

1. Ahlfors LV (1966) Complex analysis: An introduction to the theory of analytical functions of one complex variable. McGraw-Hill, New York
2. Ahlfors LV (1973) Conformal invariance: Topics in geometric function theory. McGraw-Hill, New York
3. Amoroso C, Hartmann AK, Hastings MB, Moore MA (2006) Conformal invariance and stochastic Loewner evolution processes in two-dimensional Ising spin glasses. *Phys Rev Lett* 97(4):267202. [arXiv:cond-mat/0601711](#)
4. Ash RB, Doléans CA (2000) Probability and measure theory. Academic, San Diego
5. Bak P (1999) How nature works: The science of self-organized criticality. Springer, New York
6. Bauer M, Bernard D (2002)  $SLE_\kappa$  growth processes and conformal field theory. *Phys Lett B* 543:135–138. [arXiv:math.PR/0206028](#)
7. Bauer M, Bernard D (2003) Conformal field theories of stochastic Loewner evolutions. *Comm Math Phys* 239:493–521. [arXiv:hep-th/0210015](#)
8. Bauer M, Bernard D (2003) SLE martingales and the Viasoro algebra. *Phys Lett B* 557:309–316. [arXiv:hep-th/0301064](#)
9. Bauer M, Bernard D (2004) CFTs of SLEs: The radial case. *Phys Lett B* 583:324–330. [arXiv:math-ph/0310032](#)
10. Bauer M, Bernard D (2004) Conformal transformations and the SLE partition function martingale. *Ann Henri Poincaré* 5:289–326. [arXiv:math-ph/0305061](#)
11. Bauer M, Bernard D (2004) Loewner chains. [arXiv:cond-mat/0412372](#)
12. Bauer M, Bernard D (2004) SLE, CFT and zig-zag probabilities. Proceedings of the conference 'Conformal Invariance and Random Spatial Processes', Edinburgh, July 2003. [arXiv:math-ph/0401019](#)
13. Bauer M, Bernard D (2006) 2D growth processes: SLE and Loewner chains. *Phys Rep* 432:115–221
14. Baxter RJ (1982) Exactly solved models in statistical mechanics. Academic, London
15. Beffara V (2002) The dimension of SLE curves. [arXiv:math.PR/0211322](#)

16. Beffara V (2003) Hausdorff dimensions for  $SLE_6$ . *Ann Probab* 32:2606–2629. [arXiv:math.PR/0204208](#)
17. Bernard D, Boffetta G, Celani A, Falkovich G (2006) Conformal invariance in two-dimensional turbulence. *Nat Phys* 2:124–128
18. Bernard D, Le Doussal P, Middleton AA (2006) Are domain walls in 2D spin glasses described by stochastic Loewner evolutions. [arXiv:cond-mat/0611433](#)
19. Bernard D, Boffetta G, Celani A, Falkovich G (2007) Inverse turbulent cascades and conformally invariant curves. *Phys Rev Lett* 98:024501(4). [arXiv:nlin.CD/0602017](#)
20. Binney JJ, Dowrick NJ, Fisher AJ, Newman MEJ (1992) *The theory of critical phenomena*. Clarendon, Oxford
21. Camia F, Newman CM (2003) Continuum nonsimple loops and 2D critical percolation. [arXiv:math.PR/0308122](#)
22. Cardy J (1987) Conformal invariance. In: Domb C, Lebowitz JL (eds) *Phase transitions and critical phenomena*, vol 11. Academic, London
23. Cardy J (1993) Conformal field theory comes of age. *Physics World*, June, 6:29–33
24. Cardy J (1996) *Scaling an renormalization in statistical physics*. Cambridge University Press, Cambridge
25. Cardy J (2002) Conformal invariance in percolation, self-avoiding walks and related problems. Plenary talk given at TH-2002, Paris. [arXiv:cond-mat/0209638](#)
26. Cardy J (2003) Stochastic Loewner evolution and Dyson's circular ensembles. *J Phys A* 36:L379–L408. [arXiv:math-ph/0301039](#)
27. Cardy J (2005) SLE for theoretical physicists. *Ann Phys* 318:81–118. [arXiv:cond-mat/0503313](#)
28. Cardy J (2006) The power of two dimensions. *Nat Phys* 2:67–68
29. Chaikin PM, Lubensky TC (1995) *Principles of condensed matter physics*. Cambridge University Press, Cambridge
30. de Gennes PG (1985) *Scaling concepts in polymer physics*. Cornell University Press, Ithaca
31. Duplantier B (2000) Conformally invariant fractals and potential theory. *Phys Rev Lett* 84:1363–1367. [arXiv:cond-mat/9908314](#)
32. Feder J (1988) *Fractals (physics of solids and liquids)*. Springer, New York
33. Fisch R (2007) Comment on conformal invariance and stochastic Loewner evolution processes in two-dimensional Ising spin glasses. [arXiv:0705.0046](#)
34. Fischer KH, Hertz JA (1991) *Spin glasses*. Cambridge University Press, Cambridge
35. Fortuin CM, Kasteleyn PW (1972) On the random cluster model. *Physica* 57:536–564
36. Gamsa A, Cardy J (2007) SLE in the three-state Potts model – a numerical study. [arXiv:0705.1510](#)
37. Gardiner CW (1997) *Handbook of stochastic methods*. Springer, New York
38. Gong S (1999) *The Bieberbach conjecture*. RI American 19. Mathematical Society. International, Providence
39. Gruzberg IA, Kadanoff LP (2004) The Loewner equation: Maps and shapes. *J Stat Phys* 114:1183–1198. [arXiv:cond-mat/0309292](#)
40. Jensen HJ (2000) *Self-organized criticality: Emergent complex behavior in physical and biological systems*. Cambridge University Press, Cambridge
41. Kadanoff LP (1966) Scaling laws for Ising models near  $T_c$ . *Physics* 2:263–271
42. Kadanoff LP, Berkenbusch MK (2004) Trace for the Loewner equation with singular forcing. *Nonlinearity* 17:R41–R54. [arXiv:cond-mat/0402142](#)
43. Kager W, Nienhuis B (2004) A guide to stochastic Loewner evolution and its application. *J Stat Phys* 115:1149–1229
44. Kager W, Nienhuis B, Kadanoff LP (2004) Exact solutions for loewner evolutions. *J Stat Phys* 115:805–822
45. Kauffman SA (1996) *At home in the universe: The search for the laws of self-organization and complexity*. Oxford University Press, Oxford
46. Kennedy T (2002) Monte Carlo tests of stochastic Loewner evolution predictions for the 2D self-avoiding walk. *Phys Rev Lett* 88(4):130601. [arXiv:math.PR/0112246](#)
47. Kennedy T (2004) Conformal invariance and stochastic Loewner evolution predictions for the 2D self-avoiding walk – Monte Carlo tests. *J Stat Phys* 114:51–78. [arXiv:math.PR/0207231](#)
48. Kennedy T (2005) Monte Carlo comparisons of the self-avoiding walk and SLE as parameterized curves. [arXiv:math.PR/0510604v1](#)
49. Kennedy T (2006) The length of an SLE – Monte Carlo studies. [arXiv:math.PR/0612609v1](#)
50. Kennedy T (2007) Computing the Loewner driving process of random curves in the half plane. [arXiv:math.PR/0702071v1](#)
51. Kolmogorov AN (1941) Dissipation of energy in the locally isotropic turbulence. *Dokl Akad Nauk SSSR* 30:9–13. (Reprinted in *Proc Royal Soc Lond A* 434:9–13 (1991))
52. Kraichnan RH (1967) Inertial ranges in two-dimensional turbulence. *Phys Fluids* 10:1417–1423
53. Kraichnan RH, Montgomery D (1980) Two-dimensional turbulence. *Rep Prog Phys* 43:567–619
54. Landau LD, Lifshitz EM (1959) *Theory of elasticity*. Pergamon, Oxford
55. Lawler GF (2004) *ICTP Lecture Notes Series*. 17:307–348
56. Lawler GF (2005) Conformally invariant processes in the plane. In: *Mathematical Surveys and Monographs*, vol 114. AMS, Providence
57. Lawler GF, Schramm O, Werner W (2001) Conformal invariance of planar loop-erased random walks and uniform spanning trees. *Ann Prob* 32:939–995. [arXiv:math.PR/0112234](#)
58. Lawler GF, Schramm O, Werner W (2001) The dimension of the planar Brownian frontier is 4/3. *Math Res Lett* 8:401–411. [arXiv:math.PR/00010165](#)
59. Lawler GF, Schramm O, Werner W (2001) Values of Brownian intersections exponents I: Half plane exponents. *Acta Math* 187:237–273. [arXiv:math.PR/9911084](#)
60. Lawler GF, Schramm O, Werner W (2001) Values of Brownian intersections exponents II: Plane exponents. *Acta Math* 187:275–308. [arXiv:math.PR/0003156](#)
61. Lawler GF, Schramm O, Werner W (2002) On the scaling limit of planar self-avoiding walk. *Fractal geometry and application, a jubilee of Benoit Mandelbrot*, Part 2, 339–364. *Proc. Sympos. Pure Math.*, 72, Part 2, Amer. Math. Soc., Providence, RI, 2004. [arXiv:math.PR/0204277](#)
62. Lawler GF, Schramm O, Werner W (2002) Values of Brownian intersections exponents III: Two-sided exponents. *Ann Inst Henri Poincaré* 38:109–123. [arXiv:math.PR/0005294](#)
63. Lawler GF, Schramm O, Werner W (2003) Conformal restriction: The chordal case. *J Amer Math Soc* 16:917–955. [arXiv:math.PS/0209343](#)



64. Löwner K (Loewner C) (1923) Untersuchungen über schlichte konforme Abbildungen des Einheitskreises. I Math Ann 89:103–121
65. Ma S-K (1976) Modern theory of critical phenomena. *Frontiers in Physics*, vol 46. Benjamin, Reading
66. Mackenzie D (2000) Taking the measure of the wildest dance on earth. *Science* 290:1883–1884
67. Majumbar SN (1992) Exact fractal dimension of the loop-erased self-avoiding walk in two dimensions. *Phys Rev Lett* 68:2329–2331
68. Mandelbrot B (1987) *The fractal geometry of nature*. Freeman, New York
69. Moore M (2007) private communication
70. Nicolis G (1989) *Exploring complexity: An introduction*. Freeman, New York
71. Nienhuis B (1987) Coulomb gas formulation of two-dimensional phase transitions. In: Domb C, Lebowitz JL (eds) *Phase transitions and critical phenomena*, vol 11. Academic, London
72. Pfeuty P, Toulouse G (1977) *Introduction to the renormalization group and to critical phenomena*. Wiley, New York
73. Reichl LE (1998) *A modern course in statistical physics*. Wiley, New York
74. Rohde S, Schramm O (2001) Basic properties of SLE. *Ann Math* 161:879–920. [arXiv:mathPR/0106036](https://arxiv.org/abs/mathPR/0106036)
75. Rushkin I, Oikonomou P, Kadanoff LP, Gruzberg IA (2006) Stochastic Loewner evolution driven by Levy processes. *J Stat Mech* (2006) 01:P01001. [arXiv:cond-mat/0509187](https://arxiv.org/abs/cond-mat/0509187)
76. Saleur H, Duplantier B (1987) Exact determination of the percolation hull exponent in two dimensions. *Phys Rev Lett* 58:2325–2328
77. Schramm O (2000) Scaling limit of loop-erased random walks and uniform spanning trees. *Israel J Math* 118:221–288. [arXiv:math.PR/9904022](https://arxiv.org/abs/math.PR/9904022)
78. Smirnov S (2001) Critical percolation in the plane: Conformal invariance, Cardy's formula, scaling limits. *C R Acad Sci Paris Ser I Math* 333(3):239–244
79. Smirnov S (2006) Towards conformal invariance of 2D lattice models. *Proceedings of the international congress of mathematicians (Madrid, August 22–30, 2006)*. *Eur Math Soc* 2:1421–1451
80. Smirnov S, Werner W (2001) Critical exponents for two-dimensional percolation. *Math Res Lett* 8:729–744
81. Stanley HE (1987) *Introduction to phase transitions and critical phenomena*. Oxford University Press, Oxford
82. Stauffer D, Aharony A (1994) *Introduction to percolation theory*. CRC, Boca Raton
83. Strogatz S (2003) *Sync: The emerging science of spontaneous order*. Hyperion, New York
84. Vanderzande C, Stella AL (1989) Bulk, surface and hull fractal dimension of critical Ising clusters in  $d = 2$ . *J Phys A: Math Gen* 22:L445–L451
85. Werner W (2004) Random planar curves and Schramm–Loewner evolutions. *Springer Lecture Notes in Mathematics* 1840:107–195. [arXiv:math.PR/0303354](https://arxiv.org/abs/math.PR/0303354)
86. Wilson KG, Kogut J (1974) The renormalization group and the  $\epsilon$  expansion. *Phys Rep* 12:75–199
87. Wu FY (1982) The Potts model. *Rev Mod Phys* 54:235–268
88. Zoia A, Kantor Y, Kardar M (2007) Distribution of first passage times and distances along critical curves. [arXiv:0705.1474v1](https://arxiv.org/abs/0705.1474v1) [cond-mat.stat-mech]

## Stochastic Models of Biological Processes

STEVEN S. ANDREWS<sup>1</sup>, TUAN DINH<sup>1</sup>, ADAM P. ARKIN<sup>1,2</sup>

<sup>1</sup> Physical Biosciences Division, Lawrence Berkeley National Laboratory, Berkeley, USA

<sup>2</sup> Department of Bioengineering and Howard Hughes Medical Institute, University of California, Berkeley, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Non-Spatial Stochastic Modeling](#)

[Spatial Stochastic Modeling](#)

[Future Directions](#)

[Box 1: The \*E. Coli\* Min System](#)

[Acknowledgments](#)

[Bibliography](#)

### Glossary

**Brownian dynamics** A level of detail in which each molecule is represented by a point-like particle and molecules move in response to diffusion and collisions

**Chemical Fokker–Planck equation (CFPE)** Master equation for well-mixed systems that corresponds to the chemical Langevin equation

**Chemical master equation (CME)** Master equation for the probability that the system has specific integer copy numbers for each type of chemical species; it is exact for a well-mixed system

**Chemical Langevin equation (CLE)** Approximate stochastic differential equation for well-mixed systems which is based on continuous Gaussian statistics

**Direct method** An implementation of the Gillespie algorithm

**Extrinsic noise** In genetic noise studies, expression fluctuations of a gene that arise from upstream genes or global fluctuations

**First-reaction method** An implementation of the Gillespie algorithm

**Gillespie algorithm** Exact algorithm for simulating individual trajectories of the CME

**Hybrid algorithms** Algorithms that are designed to efficiently simulate systems that have multiple timescales

**Individual-based spatial models** Models that track individual molecules as they diffuse or react



**Intrinsic noise** Expression fluctuations of a gene that arise from that particular gene

**Jump process** A process in which the system abruptly changes from one state to another

**Optimized direct method** A computationally efficient implementation of the Gillespie algorithm

**Population-based spatial models** Models that track how many molecules of each chemical species are in various spatial compartments

**Reaction channel** A possible reaction between specified reactant and product chemical species (the terminology distinguishes this meaning from an individual reaction event between single molecules)

**Reaction-diffusion equation** Deterministic partial differential equation that combines mass action reaction kinetics and normal chemical diffusion

**Reaction-diffusion master equation (RDME)** Chemical master equation that accounts for diffusion as well as reactions

**Reaction rate equation (RRE)** Deterministic ordinary differential equation for the net production rate of each chemical species from chemical reactions

**Spatial chemical Langevin equation** Chemical Langevin equation that accounts for diffusion as well as reactions

**Stochastic simulation algorithm (SSA)** Alternative term for the Gillespie algorithm

**Stoichiometric matrix ( $\nu$ )** Matrix that gives the net production of each chemical species, for each chemical reaction

**Tau-leaping method** Approximate simulation method for well-mixed systems in which molecule numbers are updated using discrete Poisson statistics

**Well-mixed hypothesis** Assumption that mixing processes occur faster than the relevant reaction processes

## Definition of the Subject

Many processes in cell biology, such as those that carry out metabolism, the cell cycle, and various types of signaling, are comprised of biochemical reaction networks. It has proven useful to study these networks using computer simulations because they allow us to quantitatively investigate hypotheses about the networks. Deterministic simulations are sufficient to predict average behaviors at the population level, but they cannot address questions about noise, random switching between stable states of the system, or the behaviors of systems with very few molecules of key species. These topics are investigated with stochastic simulations. In this article, we review the dominant types of stochastic simulation methods that are used to investigate biochemical reaction networks, as well as some of the

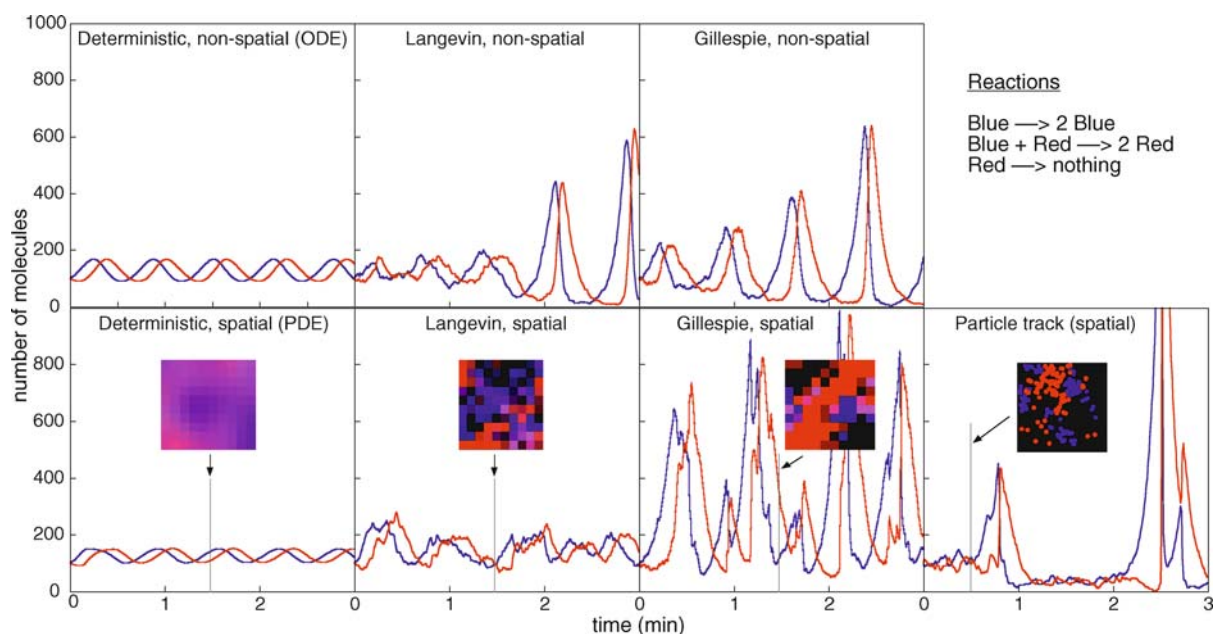
results that have been found with them. As new biological experiments continue to reveal more detail about biological systems, and as computers continue to become more powerful, researchers will increasingly turn to simulation methods that can address stochastic and spatial details.

## Introduction

Random events are ubiquitous throughout biology. Diffusion, chemical reactions, gene expression, homologous recombination, and most other fundamental biological processes are governed to a large extent by the inherently discrete and stochastic interactions of molecules [1]. In many cases, the random events that occur on very small length and time scales become averaged out when one focuses on larger length or time scales. However, there also exist many examples in which stochastic fluctuations at small scales propagate up to and then influence the system behavior at large scales. Examples range from the swimming trajectories of individual bacteria all the way up to the genetic diversity on which evolution depends.

Research on stochasticity in biochemical systems has received a great deal of attention lately, leading to many recent reviews [2,3,4,5,6,7,8,9]. One reason for its popularity is that the basic designs of many biochemical systems, such as metabolism, cell division, and chemotaxis, are becoming reasonably well understood. Starting from this understanding, researchers are delving deeper to examine the quantitative behaviors of these systems, including the roles of stochastic influences. Also, there is an increasing awareness of the importance of stochasticity in biological systems. For example, it has become clear that noisy gene expression is the rule rather than the exception [9]; this leads to important questions about non-genetic individuality and about biological robustness to gene expression noise. Thirdly, stochasticity is often investigated using computationally demanding simulations. Cheaper and faster computers, as well as improved simulation algorithms, are making it feasible for researchers to investigate more complex biochemical systems, at increasingly realistic levels of detail. Finally, and perhaps most importantly, the last ten years have witnessed incredible progress in experimental biochemical methods, some of which allow direct measurements of stochasticity on microscopic size scales. These methods include gene expression measurements [10,11], flow cytometry [12,13,14,15,16], single molecule detection methods [17,18,19,20], and applications of synthetic biology [21].

At the most fundamental level, the quantum mechanics that describe the dynamics of all physical systems are well-understood and completely deterministic [22]. It is



Stochastic Models of Biological Processes, Figure 1

Simulation results for a simple chemical oscillator using different simulation methods. The Lotka–Volterra system is shown, which shares key features with cellular oscillators such as circadian rhythms. *Insets* show the spatial distributions of molecules at the indicated times. In the *top panels*, note that stochasticity allows the system to drift to large amplitude oscillations and that the Langevin and Gillespie methods yield similar results. In the *bottom panels*, all of which were started with nearly homogeneous initial states, differences arise from the approximations: the PDE simulation has predictable oscillations due to the minimal stochasticity (which is only in the initial state); the Gillespie simulation has larger peaks than the Langevin one because it only allows integer numbers of molecules in each bin; and the particle tracking simulation shows larger and fewer bursts than does the Gillespie simulation because it accurately treats diffusion at all length scales (this difference was reduced with a spatial Gillespie simulation that used smaller subvolumes). Parameters: rate constants are  $10 \text{ min}^{-1}$ ,  $8000 \text{ nm}^3 \text{ molec}^{-1} \text{ min}^{-1}$ , and  $10 \text{ min}^{-1}$ , for the respective reactions shown in the top-right corner, systems start with 100 of each blue and red molecules, their diffusion coefficients are  $100 \text{ nm}^2 \text{ min}^{-1}$ , the volume is  $100 \text{ nm}$  high and wide by  $10 \text{ nm}$  deep, and the first three spatial simulations divide this volume into cubic subvolumes that are  $10 \text{ nm}$  on a side. This figure is reproduced from [8]

only when systems are observed that there arises unavoidable randomness, although these aspects of quantum mechanics remain murky and as close to philosophy as science. More importantly, essentially any system that is governed by nonlinear dynamics, including nearly all physical systems, rapidly becomes chaotic as the system size is increased beyond a few molecules, and thus becomes effectively unpredictable [23]. For all intents and purposes, the diffusive trajectories of individual molecules, and the probabilities of chemical reactions occurring between neighboring reactants at specific times, are fundamentally stochastic.

In the laboratory, partly by design, this stochasticity usually averages out. For systems comprised of many particles, diffusion is observed to be described well by Fick's law of diffusion and chemical reaction kinetics are described well by the deterministic reaction rate equations [24]. The situation is often different within biolog-

ical cells for several reasons: (i) the program for cellular behavior, the genome, is present at low copy number and yet each gene governs the expression of possibly thousands of proteins, (ii) the low copy numbers of many proteins and mRNA transcripts within cells make random variation of their numbers a relatively large fraction of the total, (iii) stochasticity is often amplified during sequential biochemical steps, of which an especially important case is the sequence of DNA transcription followed by mRNA translation, and (iv) because of spatial organization within cells, it often takes very few proteins in specific locations to achieve large effects on the entire cell dynamics.

This review provides a broad account of stochastic modeling of biological processes. The emphasis is placed on stochastic processes at the cellular level although much of the work presented here also applies to other scales and systems. Our goal is to briefly familiarize the reader with the mathematical forms of the most important equations,

the tools for analyzing and simulating them, and some applications for which they have been particularly successful. As shown below, the mathematics, the software implementation, and even the applications of modeling methods are often closely linked.

There are several ways to categorize stochastic biochemical modeling methods. A key designation is whether a method is spatial or non-spatial: spatial models treat spatial organization of proteins and membranes explicitly, whereas non-spatial models include an implicit assumption that mixing processes occur faster than the relevant reaction processes, which is called the *well-mixed hypothesis*. No single modeling method can efficiently capture stochastic dynamics over wide ranges of time scales, so separate methods have been developed that operate at levels of temporal detail that range from nanoseconds to hours. Hybrid simulators combine methods that operate at different timescales to allow the efficient simulation of systems that include both fast and slow processes. Whereas many modeling methods are designed solely to address the reactions in a biochemical reaction network, others also consider system boundaries, mechanics, or the multiple states that proteins can be in. Here, we present non-spatial modeling methods first, followed by spatial methods, with diversions along the way to touch on as many of the other topics as possible. Simulation results from many of the methods that are discussed are compared in Fig. 1, where it is seen that the differences can be quite significant.

## Non-Spatial Stochastic Modeling

### Deterministic Modeling and Notation

Before focusing on stochastic modeling, it is helpful to introduce the notation and some terminology by summarizing a few aspects of deterministic modeling. As applied to chemical reaction networks, deterministic modeling is based upon ordinary differential equations (ODEs) for the individual chemical reactions. Consider the generic elementary reversible reaction



where  $k_f$  and  $k_r$  are the forward and reverse reaction rates, respectively. We assume that the system is kept well-mixed so that diffusion effects can be ignored. The reaction rate equations for components  $A$ ,  $B$ , and  $C$  are the ODEs

$$\frac{d[A]}{dt} = \frac{d[B]}{dt} = -k_f[A][B] + k_r[C], \quad (2a)$$

$$\frac{d[C]}{dt} = k_f[A][B] - k_r[C]. \quad (2b)$$

More complex reaction networks are expressed analogously, with one equation for each chemical species and with terms in the equations that represent chemical reactions. From these equations, the reactions can be simulated to show how the concentrations change over time. Or, after setting the left sides of the equations to zero, they can be solved to yield the steady-state chemical concentrations. One can also investigate the dynamic or steady-state behaviors as the reaction rate parameters ( $k_f$  and  $k_r$ ), or initial concentrations, are varied [25]; this can yield phase diagrams for the reactions and additional insight.

It is helpful to generalize the rate equations given above to make them more convenient for computational or analytical work and to show their forms more clearly. First, each chemical concentration is replaced by the variable  $Z_i(t)$ , where  $i$  is an index for  $A$ ,  $B$ , or  $C$  and the time-dependence is written out explicitly. Next, the product terms in the equations are replaced by the functions  $\tilde{a}_f(\mathbf{Z}(t))$  and  $\tilde{a}_r(\mathbf{Z}(t))$  for the forward and reverse reactions, respectively; the tildes indicate that molecule quantities are given as concentrations rather than as molecule numbers. These functions are called the *reaction propensities*. Finally, the ‘+’ or ‘−’ signs show the reaction stoichiometry. They are replaced by  $\nu_{fi}$  and  $\nu_{ri}$  for the forward and reverse reactions, respectively, which are elements of the so-called *stoichiometric matrix* (throughout this review, we follow Gillespie’s notation [6]). In this example, one unit of each  $A$  and  $B$  are lost in a unit amount of forward reaction ( $\nu_{fA} = \nu_{fB} = -1$ ) as one unit of  $C$  is formed ( $\nu_{fC} = 1$ ); the  $\nu_{ri}$  values have the opposite signs. With these substitutions, the rate equations become

$$\frac{dZ_i(t)}{dt} = \nu_{fi} \tilde{a}_f(\mathbf{Z}(t)) + \nu_{ri} \tilde{a}_r(\mathbf{Z}(t)), \quad (3a)$$

$$\tilde{a}_f(\mathbf{Z}(t)) = k_f Z_A Z_B, \quad (3b)$$

$$\tilde{a}_r(\mathbf{Z}(t)) = k_r Z_C. \quad (3c)$$

These equations are trivially extended to arbitrarily large reaction networks. Consider a system with  $N$  chemical species that can react via  $M$  different *reaction channels* (a “reaction channel” is simply unambiguous terminology for a reaction between specific reactant and product chemical species). The dynamics of this system are given with the *reaction rate equation* (RRE):

$$\frac{dZ_i(t)}{dt} = \sum_{j=1}^M \nu_{ji} \tilde{a}_j(\mathbf{Z}(t)). \quad (4)$$

The reaction propensity equations are typically the products of reaction rate constants and the appropriate chemical concentrations, as shown above (Eq. (3b) and (3c)),

but they may also describe non-elementary processes such as Michaelis–Menten kinetics. It is worth noting that the state of the system, at any point in time, is fully expressed with the vector  $\mathbf{Z}(t)$ . This means that the entire trajectory of the system can be deterministically calculated from the RRE and any single  $\mathbf{Z}(t)$  snapshot.

The RRE is at the heart of many branches of quantitative biochemistry and systems biology. A great deal of metabolic theory is based on either the steady-state solutions of the RRE, or the set of steady-state solutions that are possible, given only knowledge of the stoichiometric matrix [26,27,28,29]. Studies of biochemical oscillations [25], including the cell cycle [30,31], circadian rhythms [32,33], and certain spatial patterns [34,35] are usually based on the dynamics of the deterministic RRE. Research on biochemical switches, as are found in prion diseases [36], developmental processes [37], and some protein kinase cascades [38], often focuses on the multiple steady-state solutions of the RRE [39]. Deterministic modeling has been, and still is, the conventional modeling method for most biological systems.

### The Chemical Master Equation

Although the RRE is tremendously useful, it cannot address the stochastic processes that are inherent to biochemical systems. This is because the RRE arises from a series of approximations to a more physically rigorous stochastic model of chemical reactions [40,41].

As above, we assume that diffusive processes are much faster than reactive ones [8], which allows us to ignore spatial organization (this assumption is usually valid for genetic and metabolic networks, but often invalid for signaling networks). Nevertheless, the stochasticity of diffusion plays an essential role because it makes the precise timing of individual reactions effectively random. These reactions occur in abrupt transitions, in which reactants are converted effectively instantaneously into products, making this a type of *jump process*. Also, random diffusion causes the system to rapidly lose any memory of its prior states, and thus of the sequence of reactions that led up to the current state. This independence of the system dynamics on its history, called the Markov property, implies that the probability that a specific reaction occurs depends only on the state of the system at that time [41].

Because of the random reaction timing, reactant concentrations do not follow the deterministic trajectory that is predicted by the RRE. Instead, many concentration trajectories are possible, of which a single effectively random one actually occurs. There are two primary ways to investigate the possible trajectories with computational meth-

ods. One can simultaneously track the probability of every possible outcome or one can simulate many independent stochastic trajectories and then analyze them as one would with several repetitions of an experiment. These methods are described in this and subsequent sections, respectively.

To mathematically track the probability that the system is in each possible state, it is helpful to first replace the vector of chemical concentrations that was introduced earlier,  $\mathbf{Z}(t)$ , with a vector of integer-valued molecule numbers,  $\mathbf{X}(t)$ . These are related to each other, within round-off error, through the volume of the system, which is given as  $\Omega$ ,

$$\mathbf{X}(t) \simeq \Omega \mathbf{Z}(t). \quad (5)$$

The state of the stochastic system is fully captured by  $\mathbf{X}(t)$ . The probability that the system is in state  $\mathbf{x}$  at time  $t$ , given that it started in state  $\mathbf{x}_0$  at time  $t_0$ , is written as  $P(\mathbf{x}, t | \mathbf{x}_0, t_0)$ . This probability changes over time because chemical reactions can transfer the system either into this state from other ones, or out of this state and into others. These possible transitions are combined to yield the *chemical master equation* (CME) [6,42]:

$$\frac{\partial P(\mathbf{x}, t | \mathbf{x}_0, t_0)}{\partial t} = \sum_{j=1}^M \left[ a_j(\mathbf{x} - \mathbf{v}_j) P(\mathbf{x} - \mathbf{v}_j, t | \mathbf{x}_0, t_0) - a_j(\mathbf{x}) P(\mathbf{x}, t | \mathbf{x}_0, t_0) \right]. \quad (6)$$

The sum is over the reaction channels that can occur in the system. The two terms within brackets give the rate at which the probability of being in state  $\mathbf{x}$  increases or decreases over time because of reactions into or out of state  $\mathbf{x}$ , respectively. These are proportional to the reaction propensities for the respective reactions. They are also proportional to the probability that the system was in the starting state, because the system can only leave a state if it was there in the first place.

The reaction propensity  $a_j(\mathbf{x})$ , given here without a tilde because it is for molecule numbers rather than concentrations, is a probability density:  $a_j(\mathbf{x})dt$  is the probability that exactly one reaction of type  $j$  will occur in a system in state  $\mathbf{x}$  within the next  $dt$  amount of time. This microscopic propensity function is as central to stochastic chemical kinetics as its macroscopic analog is to the reaction rate equation. However, the microscopic propensity rests on a solid microphysical basis, and has in fact been shown to have an exact solution for a well-stirred thermally-equilibrated gas-phase system [43]. For such a system, the propensity function is

$$a_j = h_j c_j, \quad (7)$$

where  $h_j$  is the number of distinct combinations of individual reactants for reaction  $j$  and  $c_j$  is the probability density for one of those reactions to occur. That is,  $c_j(t)dt$  is the probability that a randomly selected set of reactants for reaction  $j$  will collide and react in the next infinitesimal time interval  $dt$ . For a variety of reaction mechanisms, the  $c_j$  values can be calculated quite accurately from only the physical properties of the system [43].

The primary approximation made for the CME is that the reactive system is well-mixed. This implies both that there is no spatial organization and that there are no significant correlations between successive reactions (the Markov property). Examples of correlated reactions include metabolite channeling, which is the transfer of a metabolite from one enzyme to the next before it has a chance to equilibrate into the cytoplasm [44], and geminate recombinations, which are multiple bindings between molecules that bind reversibly [45]. The well-mixed statement also encompasses the assumption that the system is isothermal, which is typically the case in biological systems.

Because the CME becomes computationally intractable with any but the simplest systems, some recent work has focused on efficient solution methods. An algorithm called the finite state projection method accomplishes this by projecting a matrix form of the CME onto a smaller space [46,47]. By choosing the size of the projection space, the accuracy can be adapted to any level of precision. Less formal methods for state-space reduction of the CME have been proposed as well [48]. Another method uses a sparse grid, which can work efficiently for up to 10 proteins [49]. Work has also gone into separations of the CME into fast and slow components, as described in the section on hybrid methods [42,50,51].

### Applications of the Chemical Master Equation

The solution of the CME suffers from the so-called “curse of dimensionality” as the size of the state space, and hence the number of equations, increases exponentially with the number of chemical species involved. Except for very small and simple systems, it is extremely difficult to obtain solutions of the CME, either analytically or numerically. However, a few papers do report quite interesting results from direct simulations of the CME.

The master equation was used to investigate the dynamics of transiently denatured segments of double stranded DNA [52]. The authors derived the dynamics, formation rates, and lifetimes of these “bubbles”, which can be compared to fluorescence correlation microscopy experiments of fluorescently tagged base pairs [53]. In

another use of the CME, studies on molecular motors [54,55] demonstrate how the load-velocity curve, including rectified motion, arises from nucleotide triphosphate binding free energies. These works more fully investigate ideas on thermal ratchets that were presented previously [56]. A particularly intriguing study on the copy number control system for bacterial plasmids [57] showed that stochasticity in a regulatory portion of a system can actually decrease the stochasticity elsewhere in the system. This “stochastic-focusing” changes the behaviors of gradual-response systems towards those of threshold systems [58,59] in a manner that is analogous to the oscillation enhancement that stochastic resonance can create in oscillating systems [60]. These results contradict the widely held belief that an increase in stochasticity in one portion of a system will necessarily increase the stochasticity everywhere downstream of it. Studies of simple signal transduction motifs have shown how the predictions of the RRE can be qualitatively wrong compared to the CME treatment in that the stochastic systems might be bistable or oscillate when the deterministic system has one stable state [40,61]. Many of these studies that used the CME also used other theoretical techniques as well, which allow fruitful comparisons between the methods.

### The Gillespie Algorithm

Because of the challenges in working with the CME, it is most often investigated using a Monte Carlo approach in which individual sample trajectories are simulated. These simulations can be exact or approximate. In this context, “exact” means that if the simulation were run many times, the distribution of simulated trajectories would agree exactly with that which would be predicted by an analytical solution, were it obtainable, of the chemical master equation. Exactness implies nothing about the validity of the CME or about the limitations of the computational accuracy (such as round-off errors and imperfect pseudo-random number generators), but only that no further approximations are made beyond those that are assumed by the CME.

In 1976, Gillespie introduced an exact algorithm for simulating the CME [6,62,63] which is called the *stochastic simulation algorithm* (SSA) in his papers, but is better known as the *Gillespie algorithm*. This algorithm cycles through three portions: (i) generate the time step until the next reaction, (ii) determine which reaction that will be, and (iii) execute the reaction by advancing the time and molecule counts to reflect it. The Gillespie algorithm was introduced with two varieties, called the *direct method* and the *first-reaction method*. In the former, the time step to



the next reaction,  $\tau$ , and the reaction number,  $j$ , are chosen from the following probability distributions:

$$P(\tau) = ae^{-a\tau}, \quad (8a)$$

$$P(j) = \frac{a_j(\mathbf{X})}{a}. \quad (8b)$$

The variable  $a$  represents the summed reaction propensity,

$$a \equiv \sum_j a_j(\mathbf{X}). \quad (8c)$$

In the first-reaction method, a time step,  $\tau_j$ , is generated for each possible reaction channel. Again, these are exponentially distributed random numbers,

$$\tau_j = a_j(\mathbf{X}) e^{-a_j(\mathbf{X})\tau}. \quad (9)$$

The smallest of these time steps is chosen as the next simulation time step, while its subscript dictates the reaction channel that is executed at that time. The direct method is usually preferred because it is a little easier to program and runs slightly faster with a simple implementation. However, the latter has been favored as a basis for improvements on computational efficiency.

The exactness of the Gillespie algorithm comes at the cost of its being computationally demanding. Even if one simulation can be performed reasonably quickly, such as in a few minutes, this can still be too slow to investigate the behaviors of hundreds of mutant cells or to explore different regions of parameter space. Thus, much effort has been devoted to improving the efficiency of the Gillespie algorithm, while still maintaining exactness. These methods are all based upon either the direct method or the first-reaction method, but are carefully designed, usually with priority queues or other indexing methods, so that internal variables are recalculated as infrequently as possible [6,64,65,66,67]. Of these, it appears that the *optimized direct method* is probably best for most practical problems [66]. Because the faster methods are significantly more difficult to program than the original ones, both sets of methods are still commonly used.

The computational intensity of the Gillespie algorithm, even with more efficient implementations, makes it difficult to perform sensitivity analyses. In these analyses, one investigates the extent to which the results depend upon input parameters, which can be helpful for determining which parameters need additional experimental investigation or which are particularly important for system control. An algorithm for stochastic sensitivity analysis was recently developed and applied to biochemical reaction networks [68]. It involves the addition of just two steps to the basic loop of the Gillespie algorithm.

Another difficulty of the Gillespie algorithm, which also applies to simulations of the RRE and other algorithms presented below, is called *combinatorial explosion*. Suppose a scaffold protein has several sites with which it can bind other proteins, and suppose that each of those proteins can bind to none, one, or two phosphate groups (this is the situation for the Ste5 protein in the yeast pheromone response pathway [69]). There are clearly a tremendous number of possible binding states that the scaffold protein can be in, each of which has to be treated as a separate chemical species. Just listing all of these states is tedious, and simulating their reaction dynamics with the Gillespie algorithm is very slow. One solution is to not list every possibility when the simulation starts, but to create states when they are needed and to destroy them when they are no longer required [65]. Another option is to use an algorithm that is implemented in a program called StochSim [70,71]. Unlike the Gillespie algorithm, this one does not stochastically choose reactions to execute, but it instead chooses reactant pairs from the pool of existing molecules. A probabilistic scheme is used to determine if these reactants should be made to react with each other.

What if the system volume changes as a function of time? This might seem like an unusual concern, but it occurs during cell growth (and cell division) and it affects the reaction propensities. The necessary modifications to the Gillespie algorithm were recently derived, which are likely to be particularly useful for relatively slow processes, such as protein production from infrequently expressed genes [72].

### Applications of the Gillespie Algorithm

Models based on the Gillespie algorithm have provided critical insights into the stochastic nature of gene expression [3,73,74]. In particular, fluctuations in the rates of gene transcription are amplified at the translation stage to yield highly erratic time patterns of protein production [75]. When multiple regulatory proteins act together, or compete with each other, this randomness is amplified further because of the random sequence of protein bursts [75]. These effects were shown to stochastically switch a model of phage- $\lambda$  between the bistable lysis and lysogeny states [76], with results that are consistent with experimental ones. Stochastic gene expression is also used by many pathogenic organisms to randomly switch their surface features so they can evade host responses [77], may be used by the HIV virus to stochastically delay viral expression long enough for transformation of its activated T-cell host to a memory cell and thereby trap HIV

as a latent phage [16], can establish asymmetries that determine cell differentiation [78], and can cause circadian clocks to lose synchrony [21,79]. In fluctuating environments, stochastic gene expression can permit an isogenic bacterial population to grow faster than it would if all individuals were phenotypically homogeneous [77,80,81].

From combined modeling and experimental approaches, the dominant noise sources in the stochastic expression of a specific gene are: (i) the expression fluctuations of that particular gene, which is called the *intrinsic noise*, (ii) noise that is transmitted to it from upstream genes, and (iii) global noise that affects all genes. The latter two sources are often combined and called *extrinsic noise*, which is the total noise source that is extrinsic to that specific gene [12,74,82,83,84]. Noise arises at both the transcription and translation stages, for which the relative importance depends on the strength of the promoter and on whether prokaryotic or eukaryotic transcriptional machinery is used [14,15,75,85,86]. Direct measurements of gene expression have generally confirmed the predictions made by stochastic simulations [11,12,13,15,17,18,21,87].

Gene expression appears to be unavoidably stochastic, and this randomness is usually amplified at each stage, so how does biology function reliably amidst all the noise? This is a central topic of many papers on biological robustness [5,9,88,89,90,91], several of which use the Gillespie algorithm or other types of stochastic modeling. One answer is that many reaction network structures are inherently less susceptible to noise than others. These include ones in which the reaction rates do not depend on the number of mRNA transcripts [92], certain scale-free reaction networks [93], and networks that are designed to function near saturation [94] (analogous to binary logic). Secondly, there are several mechanisms for biological robustness to noise, including negative feedback, integral feedback [95], checkpoints, and redundancy [9]. Because gene expression noise is usually detrimental to biological function, it has been suggested that there is active selection for robustness mechanisms [96,97].

### Approximate Stochastic Methods

Because every reaction is simulated individually in the Gillespie algorithm, it is unavoidably computationally demanding, even with the algorithmic methods that have been developed to speed it up. To address this, several approximate methods have been developed.

The most accurate of these approximate methods is called the *tau-leaping* method [6,98,99,100]. In contrast to the Gillespie algorithm, the  $\tau$ -leaping method uses a simulation time step which is long enough that many individual

reactions are likely to occur during the time interval. The reaction propensities ought to change slightly as each reaction occurs to reflect the new chemical populations, although this algorithm uses the assumption that changes within a single time step are negligible. This is the sole approximation made for the  $\tau$ -leaping method. Each reaction is considered to be an independent event (a consequence of both the well-mixed hypothesis and the constant reaction propensities), so the number of reactions that occur during time step  $\tau$  for reaction channel  $j$  is a Poisson-distributed random variable; it is denoted  $k_j$  and has a mean value equal to the reaction propensity  $a_j(\mathbf{X}(t))$ . The formula used by  $\tau$ -leaping that updates the system state over one time step is

$$\mathbf{X}(t + \tau) = \mathbf{X}(t) + \sum_{j=1}^M k_j \mathbf{v}_j. \quad (10)$$

The algorithm alternates steps in which the state of the system is updated and those in which new reaction propensities are calculated.

As the time step is reduced to zero, the  $\tau$ -leaping simulation method approaches that of the Gillespie algorithm, although with more computational overhead. In the other direction, increasing the time step makes the simulation become more and more approximate. It has been suggested that  $\tau$  should be chosen by first predicting the change in molecular populations over time using deterministic methods; then, the time step is chosen so that no molecular species is likely to change its population during this time step by more than some pre-determined fraction of its total population [98]. A difficulty that can occur with  $\tau$ -leaping (which is also an issue with ODE integration), is that it is possible for a molecular species to be assigned a negative population. Several methods have been proposed to avoid this problem, some of which are also able to improve the performance of the algorithm in other ways as well [101,102,103]. Despite several papers on the development of  $\tau$ -leaping, this method has yet to be applied to novel biological problems.

Two additional approximations allow the application of many more theoretical methods. First, the vector that defines the state of the system,  $\mathbf{X}(t)$ , is allowed to take on real values as well as integer values. As one would expect, this is usually a reasonable assumption for large chemical populations and a poor assumption for low copy numbers. In particular, it can be a very poor approximation in cases where there might become no copies of a chemical species at all; an approximate value of, say, 0.01 protein copies can lead a system to entirely different outcomes than one would find with exactly 0 copies. The second

approximation is to replace the Poisson-distributed random variables that were used in the  $\tau$ -leaping algorithm with Gaussian-distributed random variables [104]. The effect of this change again decreases as the copy numbers of chemical species are increased. It also decreases as longer time steps are used because that leads to more individual chemical reactions per time step. These approximations allow the updating equation of the  $\tau$ -leaping algorithm to be replaced by a stochastic differential equation called the *chemical Langevin equation* [105,106,107] (CLE),

$$\frac{dX_i(t)}{dt} = \sum_{j=1}^M v_{ji} a_j(\mathbf{X}(t)) + \sum_{j=1}^M \Gamma_j(t) v_{ji} \sqrt{a_j(\mathbf{X}(t))}. \quad (11)$$

The first term is simply the reaction rate equation that is given above (Eq. (4)), but for molecule counts rather than concentrations. The second term adds Gaussian noise to the deterministic result, where  $\Gamma_j(t)$  represents a temporally uncorrelated, statistically independent Gaussian white noise with mean 0 and variance 1. In other words, the integral of  $\Gamma_j(t)$  is a one-dimensional continuous random walk. The chemical Langevin equation describes a continuous Markov process which is an approximation of the jump Markov process that underlies the chemical master equation. Simulations with the CLE, which are based on an equation that is quite similar to Eq. (11) [106], yield single stochastic trajectories of the system state, much like simulations with the Gillespie algorithm or the  $\tau$ -leaping method.

Alternatively, instead of following a single system as it moves along one of its many possible stochastic trajectories, it is possible to focus on a single portion of the state space to see how likely it is that the system will be in this region of state space as a function of time. This latter picture is described by *chemical Fokker-Planck equation* [41,105,107] (CFPE),

$$\begin{aligned} \frac{\partial P(\mathbf{x}, t \mid \mathbf{x}_0, t_0)}{\partial t} = & - \sum_{i=1}^N \frac{\partial}{\partial x_i} \left[ \left( \sum_{j=1}^M v_{ji} a_j(\mathbf{x}) \right) P(\mathbf{x}, t \mid \mathbf{x}_0, t_0) \right] \\ & + \frac{1}{2} \sum_{i=1}^N \frac{\partial^2}{\partial x_i^2} \left[ \left( \sum_{j=1}^M v_{ji}^2 a_j(\mathbf{x}) \right) P(\mathbf{x}, t \mid \mathbf{x}_0, t_0) \right] \\ & + \sum_{\substack{i,i'=1 \\ i < i'}}^N \frac{\partial^2}{\partial x_i \partial x_{i'}} \left[ \left( \sum_{j=1}^M v_{ji} v_{ji'} a_j(\mathbf{x}) \right) P(\mathbf{x}, t \mid \mathbf{x}_0, t_0) \right]. \end{aligned} \quad (12)$$

The first term, often called the drift term, arises from the deterministic behavior of the system. The latter two terms, collectively called the diffusion term, represent the stochastic deviations away from deterministic behavior. Because the CFPE represents continuous processes, it is significantly more analytically tractable than the chemical master equation.

To compute the probabilities of possible system behaviors using the CFPE, state space is usually discretized into a grid and then the CFPE is integrated using standard numerical methods [108,109]. This analysis method is similar to that employed for the CME, but is usually less computationally demanding because the discretized state space is typically significantly coarser. Nevertheless, the dimensionality of state space still increases exponentially with additional chemical species, so the CFPE still suffers from the curse of dimensionality.

### Applications of Approximate Stochastic Methods

Perhaps because stochastic simulations are still a relatively new field of study, many studies with the CLE and CFPE focus more on the mathematical techniques than on the biological applications [109,110,111,112,113]. The approximate CLE and CFPE have been shown to yield results that are in good agreement with exact simulations for a reversible isomerization reaction, even with very few molecules [107].

The CLE was used to analytically investigate the role of noise-induced phenomena in enzymatic futile cycles, which is a motif that is common to many biochemical networks [61]. The analysis indicated that the presence of external noise is sufficient to induce switching bistability in the system, a phenomenon that is often attributed to feedback loops [25]. In combination with experimental data, the CLE was also used to show that translational efficiency is the predominant source of intracellular noise for a single-gene system [15]. The Fokker-Planck equation has been used to model cell growth [111,112] and cell migration [113,114]. Of particular interest, the Fokker-Planck equation has provided a convenient framework to describe the behaviors of molecular motors [109]. A motor protein is approximated as a diffusion particle in a periodic asymmetric free-energy surface. Under the input of chemical energy, the motor switches stochastically between different potentials that describe distinct biochemical states of the motor. The model has been used to explain key experimental observations for molecular motors, most notably for the  $F_1F_0$ -ATPase system [115] and a bacterial flagellar motor [109,116].

## Hybrid Algorithms

Systems that involve multiple time scales provide major simulation challenges. If the fast time scale is simulated with high precision, then the simulation takes too long for the dynamics of the slow one to be observed with any reasonable efficiency. On the other hand, if the simulation time steps are optimized for the slow timescale, then they are too long for the fast reactions and numerical errors become problematic. In the language of differential equations, these are stiff systems which require special solution techniques. For stochastic simulations with multiple timescales, several methods have been developed recently.

One class of hybrid methods focuses on new mathematics to allow approximations of the Gillespie algorithm, or related algorithms, to function with reasonable accuracy over a wide range of timescales [42,50,51,117,118,119,120,121]. The other class generally involves the coupling of multiple simulators, usually including ODE, Langevin, and/or Gillespie; the high-population molecular species are simulated with less stochastic detail and the low-population species are simulated with more stochastic detail [122,123].

## Spatial Stochastic Modeling

Most biological systems are highly organized. For example, *Escherichia coli* bacteria have helical cytoskeletons, polar-localized proteins, centrally positioned chromosomes, and elaborate flagellar motor complexes. Eukaryotes are even more organized, with elaborate organelles, microtubules and other complex cytoskeletal elements, motor proteins that shuttle back and forth, and carefully controlled cell shapes. Even phages display remarkable order in the way the DNA is packed into the outer shell. Where does this order come from? And how does this order influence the biochemical reaction network? These questions are being investigated with new imaging experiments [124,212] and with new computer simulation methods that can account for spatial heterogeneity. These spatial simulation methods are the focus of this section.

Spatial simulations have been used to investigate a wide variety of topics. These include: morphogen gradients across *Drosophila* and *Xenopus* oocytes [125,126,127], the *Escherichia coli* cell division plane localization system [128,129,130,131,132,133,134] (see Sect. “Box 1: The *E. Coli* Min System”), intracellular signaling [135,136,137,138], and rebinding of ligands to receptor complexes [139,140,141].

As described above, many successful biochemical models do not account for spatial heterogeneity; in fact, non-spatial models are in the vast majority. Typically,

non-spatial models get away with ignoring space because they model dynamics that occur more slowly than the time it takes for a molecule to diffuse across a cell, because they investigate processes that are not intrinsically spatial, and because they do not demand high quantitative accuracy. As the tools are becoming available, including both fast computers and new software algorithms [45,142,143,144], the interest in including spatial detail is increasing. These spatial models can be either deterministic or stochastic, of which our primary focus is on the latter ones.

As with the non-spatial methods that are described above, stochastic effects in spatial models arise from the discreteness of molecules. This leads to fluctuations in the numbers of molecules, which are typically on the order of the square root of the number of molecules in the appropriate characteristic volume (near steady-state and equilibrium points, but frequently greater near critical points). In spatial models, the characteristic length scale is no longer the size of the entire system but is dictated by the length scale of the spatial heterogeneity. With the shorter length scale, the characteristic volume size is reduced, fewer molecules are in these volumes, and stochastic effects increase. Thus, stochastic simulations can be required for spatial models, even if they were not needed for the corresponding non-spatial model. There are also other good reasons to model stochastic effects in spatial simulations. Many spatial phenomena, such as noise that arises from ligand rebinding [141], cannot be adequately treated without considering the detailed molecular interactions. Finally, a model is only as good as its weakest aspect. If one increases the accuracy of a model in one way, such as by accurately treating either space or stochastics, then the benefits may not be realized until the other aspect is addressed as well.

Schemes for investigating a chemical system with spatial and stochastic detail can be classified by whether they consider molecules within populations or as individuals. In the former case, space is divided it into small subvolumes, whereas in the latter, space is continuous. These classes are described in detail below. Another approach is lattice-based methods [145,146,147,148,149]. However, we do not discuss them here because they are rarely used for quantitative modeling. Furthermore, the underlying lattice geometry usually affects the results, thus making them less realistic.

## Population-Based Spatial Models

In a top-down approach towards spatial modeling, one starts with a simple, deterministic, macroscopic description and then adds successive layers of detail. In this case,

the natural starting point is with the standard textbook descriptions of chemical reactions and diffusion [24]. Reactions are described with mass action reaction kinetics expressed with the reaction rate equation that was discussed above (Eq. (4)). Diffusion is described with the diffusion equation [24], also called Fick's second law of diffusion, which is

$$\frac{\partial Z_i(\mathbf{r}, t)}{\partial t} = D_i \nabla^2 Z_i(\mathbf{r}, t). \quad (13)$$

In an extension of the definition given before,  $Z_i(\mathbf{r}, t)$  is the concentration of component  $i$  at the 3-dimensional position  $\mathbf{r}$  and time  $t$ .  $D_i$  is the diffusion coefficient for component  $i$ .

Because reactions and diffusion occur simultaneously, the respective equations are combined to express the simultaneous effects of both processes to yield the *reaction-diffusion equation*,

$$\frac{\partial Z_i(\mathbf{r}, t)}{\partial t} = \sum_{j=1}^M v_{ji} \tilde{a}_j(\mathbf{Z}(\mathbf{r}, t)) + D_i \nabla^2 Z_i(\mathbf{r}, t). \quad (14)$$

This partial differential equation (PDE) underlies a great deal of theory on chemical and biological pattern formation [126,136,150,151]. The Virtual Cell computer program [152] is general-purpose software that simulates the reaction-diffusion equation. It has been used primarily to explore spatial effects in intracellular signaling [153,154,155].

The reaction-diffusion equation is deterministic, so it captures neither the discreteness of reaction events nor the Brownian motion processes that underlie diffusion. It is possible to add this stochasticity directly into the deterministic theory but that would create a set of coupled stochastic scalar field equations, which would be extraordinarily complicated. Neither the deterministic nor the stochastic PDEs are tractable to work with analytically for any but the very simplest systems except, perhaps, in steady-state. Thus, most analysis is either computational or approximate.

In most such analyses, the equations are first simplified by dividing the system volume into an array of small cubic subvolumes, each with width  $l$ . This spatial discretization changes the diffusion portion of the reaction-diffusion equation into a discrete form:

$$\begin{aligned} \frac{dZ_{i,k}(t)}{dt} = & \sum_{j=1}^M v_{ji} \tilde{a}_j(\mathbf{Z}_k(t)) + \frac{D_i}{l^2} \sum_{k'} [Z_{i,k'}(t) - Z_{i,k}(t)]. \end{aligned} \quad (15)$$

The index  $k$  denotes the subvolume number, much as  $\mathbf{r}$  represented the spatial location. The latter summation in this discrete reaction-diffusion equation extends over all nearest neighbors of subvolume  $k$ , denoted by  $k'$ . Because the description of space was changed from continuous states to discrete states, much like the discrete kinds of molecules that are labeled by the index  $i$ , diffusion is now formally identical to reactions. The "reaction rate constant" for diffusion [156] between one subvolume and its neighbor is  $D_i/l^2$ . Because of this mathematical equivalence, much of the following discussion on the stochastic simulation of the reaction-diffusion equation parallels the discussion presented earlier on non-spatial stochastic simulations.

The first spatial stochastic equation that we present is the one that accounts for the least detail. It is the *spatial chemical Langevin equation*, which results from adding white Gaussian noise to the discrete reaction-diffusion equation. It is

$$\begin{aligned} \frac{dZ_{i,k}(t)}{dt} = & \sum_{j=1}^M v_{ji} \left[ \tilde{a}_j(\mathbf{Z}_k(t)) + \Gamma_j(t) \sqrt{\tilde{a}_j(\mathbf{Z}_k(t))} \right] \\ & + \sum_{k'} \left\{ \frac{D_i}{l^2} [Z_{i,k'}(t) - Z_{i,k}(t)] \right. \\ & \left. + \Gamma_{k'}(t) \sqrt{\frac{D_i}{l^2}} \left[ \sqrt{Z_{i,k'}(t)} - \sqrt{Z_{i,k}(t)} \right] \right\}. \end{aligned} \quad (16)$$

In an extension to what was presented before,  $\Gamma_j(t)$  and  $\Gamma_{k'}(t)$  represent temporally uncorrelated, statistically independent Gaussian white noises [106]. This is a specific example of the more general multivariate Langevin equation; it, and the multivariate Fokker-Planck equation, have been explored in depth [41,105]. However, the more specific spatial chemical Langevin equation has essentially never been used, investigated mathematically, or simulated. The sole exception that we are aware of was its simulation for a figure for a tutorial article [8] (those results are reproduced in Fig. 1).

The spatial chemical Langevin equation captures stochasticity reasonably accurately for systems in which there are many molecules per subvolume but not for those with few molecules per subvolume. Errors arise both because Gaussian white noise is the incorrect fluctuation distribution [100,104] and because it treats molecule amounts as continuously variable quantities. These are addressed by moving to the next level of detail in which the continuous molecular concentrations are replaced by discrete numbers of molecules. This changes the temporally



continuous reaction and diffusion processes to stochastic jump processes. The *reaction-diffusion master equation* [156,157,158,159] (RDME) describes the time dependence of the system at this level of description. It is

$$\begin{aligned} \frac{dP(\mathbf{x}, t)}{dt} = & \sum_{j=1}^M \left[ a_j(\mathbf{x} - \mathbf{v}_j) P(\mathbf{x} - \mathbf{v}_j, t) - a_j(\mathbf{x}) P(\mathbf{x}, t) \right] \\ & + \sum_{i=1}^N \frac{D_i}{l^2} \sum_{k,k'} \left[ (X_{i,k} + 1) P(\dots, X_{i,k} + 1, \right. \\ & \left. X_{i,k'} - 1, \dots, t) - X_{i,k} P(\mathbf{x}, t) \right]. \end{aligned} \quad (17)$$

$P(\mathbf{x}, t)$  is the probability that the system is in state  $\mathbf{X} = \mathbf{x}$  at time  $t$ ,  $X_{i,k}$  is the number of molecules of type  $i$  in subvolume  $k$ ,  $\mathbf{X}$  is the vector of all  $X_{i,k}$  values, and  $a_j$  is the propensity of reaction  $j$ .

The RDME expresses as much detail as is possible through these successive improvements of the reaction-diffusion equation. It is tempting to think of it as the fundamental equation for reactions and diffusion, and thus the basis for a statistical theory of chemistry. In fact, it is sufficiently accurate for most systems, but it nevertheless involves approximations that can be important in some situations. Firstly, neither of the starting equations, which are mass action kinetics and Fickian diffusion, are completely accurate even for very large systems. Mass action kinetics does not address the increased reaction rates that occur on extremely short time scales, which arise from reduced spatial correlations [160,161]. Nor does it address the geminate recombinations that can occur between the products of a dissociation reaction [162,163]. The diffusion equation is usually quite accurate for dilute solutions but fails for highly crowded ones [164,165], including most biological systems [166]. Secondly, the discretization of space into small subvolumes can also lead to inaccuracies, or exacerbate the inaccuracies just mentioned. The subvolume sizes must not be so small that they impinge on the microscopic details of the reaction or diffusion processes. This means that they need to be significantly larger than single molecules and larger than the mean free path lengths of diffusion [156,167]. Conversely, the subvolumes must not be so large that there would be appreciable concentration gradients across them. This means that the subvolume width needs to be less than the reactant correlation length. The correlation length is hard to predict but is at least as large as the average distance that a reactant travels before it reacts, called the reactive mean free path [156,167].

The RDME is even more intractable than the non-spatial chemical master equation because of the addition of spatial states and the many transitions that can occur between the spatial states. These additional states and transitions also make stochastic simulations of the RDME with Gillespie's direct algorithm extremely slow [157]. Several faster algorithms have been developed to address this problem. The "next reaction method" of Gibson and Bruck [64] was adapted to spatial simulations [168], and then further improved, to yield the "next subvolume method" [132,142]. Also, a fast version of the direct method [169] has been developed for spatial simulations [167]. All of these methods yield exactly the same results as Gillespie's original methods [62,63] but use carefully optimized data structures to minimize the number of computations.

A separate challenge with simulating the RDME concerns the cubical subvolumes into which space was discretized. Biological systems rarely have square corners, so the basic theory requires adaptation to account for realistic boundaries. In one approach, the mathematics was developed for dividing boundary subvolumes into two separate portions [170]. Using another approach, the theory was developed for curved surfaces, which was implemented in the MesoRD program [132,171]. Although it has not been developed yet, it has been proposed that automatic mesh refinement could simultaneously account for complex boundaries and lead to significant computational efficiencies [167].

Along with simulations of the *E. coli* Min system, presented in Sect. "Box 1: The *E. Coli* Min System", population-based spatial stochastic models have been used for a variety of test systems. In the first implementation of a spatial Gillespie algorithm, Stundzia and Lumsden used a one-dimensional simulation to demonstrate stochastic calcium wave propagation [157]. Elf and Ehrenberg showed that spatial and stochastic effects can cause an intrinsically bistable system to lose its global hysteresis through the formation of spatial domains [142]. In a third study, Isaacson and Peskin demonstrated their method for simulating porous boundaries with a model that includes transcription, translation, and nuclear membrane transport [170].

### Individual-Based Spatial Models

In a bottom-up approach to spatial modeling, one starts with a very detailed consideration and then makes successive approximations. A convenient place to start is by considering every individual molecule in the system, along with some of the molecular structures. The motions of

these molecules are governed by physical forces including steric repulsion, bond mechanics, and electrostatics. The simulation of the motions that result from these forces is called molecular dynamics [172]. Molecular dynamics can yield very accurate results but is so computationally intensive that it is rarely used for more than hundreds of cubic nanometers of volume or more than tens of nanoseconds of time. These size and time scales are too confining for studying biochemical reaction networks, so approximations are made.

At the Smoluchowski level of detail, all solvent molecules are ignored, solute molecules are treated as spheres, diffusion proceeds stochastically, and molecular rotation, molecular momentum, and long-range intermolecular forces are all ignored. This is a vast simplification, but is often valid. It is usually reasonably accurate for size scales that are larger than a few nanometers and for timescales that are longer than a few nanoseconds, constraints that are acceptable for an enormous range of chemical and biological phenomena.

For diffusion at the Smoluchowski level of detail, the effects of solvent-solute interactions on the solute motion are approximated by assuming that solute molecules diffuse with mathematically ideal Brownian motion [173,174]. This is a key approximation that replaces the deterministic molecular motions that result from solvent collisions with stochastic trajectories. It is often the only source of stochasticity in the theory, or in simulations that derive from this individual-based approach. More precisely, the position of molecule  $i$  at time  $t$  is given with the probability density  $p_i(\mathbf{r}, t)$ , which evolves over time according to the master equation

$$\frac{\partial p_i(\mathbf{r}, t)}{\partial t} = D_i \nabla^2 p_i(\mathbf{r}, t). \quad (18)$$

This equation is nearly identical to the diffusion equation, given above (Eq. (13)), differing only in the definitions of the variables and the interpretation. Now, it is not a population of molecules that diffuse, but the positional probability density for a single molecule.

Because it is so simple, the diffusion master equation is analytically tractable, in contrast to the other master equations that were discussed. One result is an entire body of analytical theory on diffusion-influenced reactions [160,175]. Nevertheless, it too becomes unmanageable for systems that have several interacting molecules, so it is simulated with a technique called *Brownian dynamics* [176,177,178,179,180]. In this method, molecules have well-defined point-like positions which are updated at each simulation time step using random displacements. The displacements are chosen by solving the diffusion

master equation for molecule  $i$ , which is taken to be at the well-defined position  $\mathbf{r}_0$  at time  $t_0$ . One simulation time step later, at time  $t_0 + \Delta t$ , the probability density for the molecule's position is found to be a 3-dimensional Gaussian density that is centered at  $\mathbf{r}_0$ ,

$$p_i(\mathbf{r}, \Delta t) = \frac{1}{s_i^3 (2\pi)^{3/2}} \exp \left[ -\frac{(\mathbf{r} - \mathbf{r}_0)^2}{2s_i^2} \right]. \quad (19a)$$

The standard deviation of this Gaussian, called the root mean square step length, is

$$s_i = \sqrt{2D_i \Delta t}. \quad (19b)$$

Brownian dynamics simulations provide accuracy that is below that of molecular dynamics, but still captures single molecule behavior.

Brownian dynamics has been used extensively for examining the rates of diffusion-influenced chemical reactions in solution [139,177,179,180,181,182] and for the rates of binding between ligands and receptor arrays [139, 141,178,183,184]. In these studies, simulated molecules diffuse in solution; at the moment that a reactant pair, or a ligand and its cognate receptor, come into contact, they undergo a chemical reaction. While diffusing, intermolecular forces are often ignored, although some studies account for these interactions as well [185,186].

To achieve the necessary level of detail, Brownian dynamics simulations usually use very short simulation time steps, often on the order of picoseconds [139]. Adaptive time steps, such that time steps are long when reactants are widely separated and short when they are close, can speed simulations up by several orders of magnitude, but are easy to implement only if there is just one diffusing particle present in the simulation volume [141]. A more sophisticated method that has the same general goal of computational efficiency is called Green's function reaction dynamics [143,187,188] (GFRD). In GFRD, which works with any number of molecules, the system is inspected to see how soon the next molecular collision or reaction could occur. The system is then advanced to that time using a single simulation time step, the event is executed if appropriate, and the cycle repeats. Yet another method, used in a program written by one of us (SSA) called Smoldyn, achieves computational efficiency by modifying the effective radii of simulated molecules so that the same reaction rate is achieved with long simulation time steps as with short ones [45,189,190]. This method does not achieve the same spatial or temporal precision as classical Brownian dynamics or GFRD, but the level of detail is still more than adequate for most biological applications and has been

shown to be indistinguishable from more accurate simulations in many cases [191].

Technically, all of these algorithms execute Brownian dynamics. However, the term “Brownian dynamics” is typically used to describe highly detailed studies in which reaction rates, rebinding dynamics, or similar phenomena are found from fundamental molecular properties such as molecular radii and intermolecular forces. In contrast, GFRD and the methods used in Smoldyn are more often used to determine system-level behaviors from known or estimated reaction rates. These are more often called particle-based stochastic simulation methods [192].

MCell is another program that performs particle-based stochastic simulations [144]. Unlike the others, it cannot simulate reactions that occur in free solution, but instead only treats reactions at surfaces. Despite the decrease of versatility, it is still useful for studying a wide variety of biological phenomena [193,194,195]; in particular, it was developed to investigate the neuromuscular junction [196,197,198]. In MCell, surface-bound receptors are not modeled as single molecules as they would be in Smoldyn or GFRD methods, but as a uniform binding probability that applies to an entire surface tile. This decreases the spatial resolution some, but increases the computational efficiency.

### Future Directions

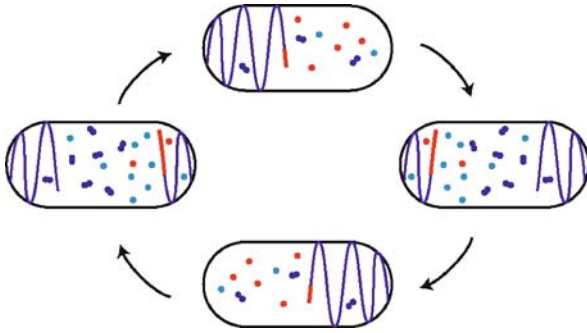
The classic advice of using the right tool for the job is as true in biochemical modeling as it is elsewhere. Several modeling tools have been presented here. Deterministic ordinary differential equation models are simple, easy to use, and can be analyzed with many powerful theoretical and analytical methods. They are the right tool for systems that can be treated as being well-mixed and that are both large enough and sufficiently far from critical points that stochastic effects are unimportant. In contrast, systems that include low copy numbers of important components, and/or that can be triggered by random events, require stochastic modeling methods for their investigation. These include integration of the chemical master equation and random sampling of the stochastic trajectories using the Gillespie algorithm, both of which are exact methods. Approximate methods include  $\tau$ -leaping stochastic simulations, integration of the chemical Fokker–Planck equation, and sampling with the chemical Langevin equation. Of these, the Gillespie algorithm has proven to be the most popular. Finally, if the system cannot be considered to be well-mixed, then yet different tools are needed. These include spatial variants of the same list of simulation methods, including partial differ-

ential equations for deterministic simulations and a spatial Gillespie algorithm for stochastic simulations. Particle-tracking simulation methods allow an even greater degree of detail.

In general, more detailed simulation methods yield more accurate results and are based more closely on underlying processes and less on phenomenological descriptions. However, they are also more computationally intensive and require more model parameters. This parametrization poses a significant problem for current models because the necessary quantitative experimental data are typically only marginally adequate or are completely non-existent. For an ODE model, it is sometimes possible to address this problem by exploring model behaviors over wide ranges of parameter space, from which one can draw phase diagrams that graphically depict how the model behaves for different parameter choices. From this, one can sometimes constrain parameters or gain additional insight into the model; for example, Tyson showed how two enzyme concentrations can be used to regulate the cell cycle, bringing an oocyte from metaphase arrest to autonomous oscillations, and on to growth-controlled cell division [30]. Because of the computational demands of spatial and stochastic models, as well as the richer behavior possibilities, it is much more difficult to explore parameter space with these more complex models. Thus, much work is needed on this topic.

More generally, the mathematical infrastructure for designing and interpreting stochastic models lags far behind that for non-spatial deterministic models. This poses some challenges for theorists. For example, what new theories and graphical tools will help scientists gain intuition into the dynamics of stochastic systems? and what are the controlling elements of stochastic systems? The theory is even more unexplored when spatial organization is considered as well. Nevertheless, spatial considerations are essential because no biological life has been found that is well-mixed; instead, a tremendous amount of biochemical activity involves membranes, polymers, protein scaffolds, large multimeric complexes, and other spatial structures. Theories that address these topics will not be as elegant as those that focus on the chemical master equation, but biology is not always elegant either.

Although research on stochastic modeling of biochemistry grew slowly from the 1950s to the 1990s, the pace has accelerated dramatically during the last 10 to 15 years. This acceleration will likely continue for many more years, in response to the faster computers that become available every year and to the ever-increasing complexity of biochemical data. With this growth, stochastic modeling may open up entire new ways to understand cell biology.



**Stochastic Models of Biological Processes, Figure 2**

Diagram of the *E. coli* Min system, which is used to position the cell division plane at the cell center. Dots represent cytoplasmic proteins, while curved lines represent helical membrane-bound protein polymers. Colors identify the proteins: light blue for MinD bound to ADP, dark blue for MinD protein bound to ATP, and red for MinE. The system dynamics are summarized in the text of Sect. "Box 1: The *E. Coli* Min System"

### Box 1: The *E. Coli* Min System

The *E. coli* Min system has served as a proving ground for spatial stochastic simulation methods. The Min system is used by *E. coli*, in conjunction with other systems, to position the cell division plane accurately at the cell center [131]. The system is comprised of the proteins MinC, MinD and MinE, which oscillate back and forth across the cell, from one pole to the other, with a period of about 40s (Fig. 2). Of these, only MinD and MinE are required for the oscillation, making this a relatively simple system that exhibits remarkably interesting dynamics. Cytoplasmic MinD proteins bind ATP, dimerize, and polymerize on the inside of the cell membrane to form long helical structures that extend outwards from one of the two cell poles. When MinE binds to the cell-center end of a MinD polymer, it activates ATP hydrolysis which depolymerizes the terminal subunit. As MinE progressively disassembles a MinD polymer at one end of the cell, it reassembles again from the opposite pole to start the next oscillation cycle. The oscillating Min proteins continually inhibit cell division plane formation near the poles using MinC, which colocalizes with MinD, thus only permitting cell division at the cell center.

This system was explored for several years with deterministic reaction-diffusion models [128,199,200]. One of these models, by Howard, Rutenberg, and de Vet [199], was also explored by the same group using a one-dimensional population-based stochastic method [129] (it uses discrete particle numbers and fixed time steps, thus conceptually placing it between the spatial Langevin and spatial Gillespie methods). The authors found that

stochastic effects were essential for generating oscillations in some parameter regimes, in a spatial version of stochastic resonance [60,201]. These models helped direct new experiments [202,203,204,205] that clarified the processes of the system.

Building on the prior models and the new experimental data, Huang, Meir, and Wingreen [130] developed a new reaction-diffusion model that was more closely connected with the biology than were previous models and that accounted for several mutant phenotypes. This model became the basis of several stochastic simulations. The spatial Gillespie method was employed by Fange and Elf [132] using their MesoRD program. They showed that a stochastic model can account for a "spotty" phenotype and for oscillations in spherical mutant cells, neither of which can be explained by the deterministic model. The MCell particle-tracking program was used by Kerr and coworkers [133] to show that the Min system alone is insufficient to center the cell division plane with high accuracy.

Yet unexplained with these simulations were convincing experimental results that MinD forms polymers on the cell membrane [205,206,207]. These were explored with another particle-tracking model [208], using a method based on Smoluchowski dynamics [45]. Although this group did simulate spontaneous polymer formation, they observed many randomly oriented short filaments, in contrast to the few helical polymers that are observed experimentally. This inherent difficulty with the reaction-diffusion model [209], whether deterministic or stochastic, has led to several studies that have focused specifically on the polymer dynamics and shapes [134,210,211].

The *E. coli* Min system is already well on its way to becoming the prototypical system for studying spatial biochemical dynamics, much as *E. coli* chemotaxis has become the prototypical system for investigating bacterial signaling.

### Acknowledgments

This work was funded by the US Department of Energy.

### Bibliography

- McQuarrie DA (1967) Stochastic approach to chemical kinetics. *J Appl Probab* 4:413–478
- Turner TE, Schnell S, Burrage K (2004) Stochastic approaches for modelling in vivo reactions. *Comp Biol Chem* 28:165–178
- Raser JM, O'Shea EK (2005) Noise in gene expression: Origins, consequences, and control. *Science* 309:2010–2013
- Samoilov MS, Price G, Arkin AP (2006) From fluctuations to phenotypes: The physiology of noise. *Sci STKE* 2006:re17

5. Rao CV, Wolf DM, Arkin AP (2002) Control, exploitation, and tolerance of intracellular noise. *Nature* 420:231–237
6. Gillespie DT (2007) Stochastic simulation of chemical kinetics. *Ann Rev Phys Chem* 58:35–55
7. Wolf DM, Arkin AP (2003) Motifs, modules and games in bacteria. *Curr Opin Microbiol* 6:125–134
8. Andrews SS, Arkin AP (2006) Simulating cell biology. *Curr Biol* 16:R523–R527
9. McAdams HH, Arkin A (1999) It's a noisy business! Genetic regulation at the nanomolar scale. *Trends Genet* 15:65–69
10. Singer RH, Lawrence DS, Ovryn B, Condeelis J (2005) Imaging of gene expression in living cells and tissues. *Biomed J Optics* 10:051406
11. Levsky JM, Shenoy SM, Pezo RC, Singer RH (2002) Single-cell gene expression profiling. *Science* 297:836–840
12. Elowitz MB, Levine AJ, Siggia ED, Swain PS (2002) Stochastic gene expression in a single cell. *Science* 297:1183–1186
13. Raser JM, O'Shea EK (2004) Control of stochasticity in eukaryotic gene expression. *Science* 304:1811–1814
14. Blake WJ, Kaern M, Cantor CR, Collins JJ (2003) Noise in eukaryotic gene expression. *Nature* 422:633–637
15. Ozbudak EM, Thattai M, Kurtser I, Grossman AD, van Oudenaarden A (2002) Regulation of noise in the expression of a single gene. *Nature Genet* 31:69–73
16. Weinberger LS, Burnett JC, Toettcher JE, Arkin AP, Schaffer DV (2005) Stochastic gene expression in a lentiviral positive-feedback loop: HIV-1 Tat fluctuations drive phenotypic diversity. *Cell* 122:169–182
17. Cai L, Friedman N, Xie XS (2006) Stochastic protein expression in individual cells at the single molecule level. *Nature* 440:358–362
18. Yu J, Xiao J, Ren X, Lao K, Xie XS (2006) Probing gene expression in live cells, one protein molecule at a time. *Science* 311:1600–1603
19. Golding I, Cox EC (2006) Protein synthesis molecule by molecule. *Genome Biol* 7:212
20. Fusco D, Accornero N, Lavoie B, Shenoy SM, Blanchard J-M, Singer RH, Bertrand E (2003) Single mRNA molecules demonstrate probabilistic movement in living mammalian cells. *Curr Biol* 13:161–167
21. Elowitz MB, Leibler S (2000) A synthetic oscillatory network of transcriptional regulators. *Nature* 403:335–338
22. Sakurai JJ (1994) *Modern Quantum Mechanics*. Addison-Wesley, Boston
23. Strogatz SH (1994) *Nonlinear Dynamics and Chaos*. Westview Press, Cambridge
24. Atkins PW (1986) *Physical Chemistry*. Freeman, New York
25. Tyson JJ, Chen KC, Novak B (2003) Sniffers, buzzers, toggles, and blinkers: dynamics of regulatory and signaling pathways in the cell. *Curr Opin Cell Biol* 15:221–231
26. Covert MW, Schilling CH, Famili I, Edwards JS, Goryanin II, Selkov E, Palsson BO (2001) Metabolic modeling of microbial strains in silico. *Trends Biochem Sci* 26:179–186
27. Varma A, Palsson BO (1994) Metabolic flux balancing: Basic concepts, scientific and practical use. *Nature Biotech* 12:994–998
28. Kauffman KJ, Prakash P, Edwards JS (2003) Advances in flux balance analysis. *Curr Opin Biotech* 14:491–496
29. Fell D (1997) *Understanding the Control of Metabolism*. Portland Press, London
30. Tyson JJ (1991) Modeling the cell division cycle: cdc2 and cyclin interactions. *Proc Natl Acad Sci USA* 88:7328–7332
31. Chen KC, Calzone L, Csikasz-Nagy A, Cross FR, Novak B, Tyson JJ (2004) Integrative analysis of cell cycle control in budding yeast. *Mol Biol Cell* 15:3841–3862
32. Goldbeter A (2002) Computational approaches to cellular rhythms. *Nature* 420:238–245
33. van Zon JS, Lubensky DK, Altena PRH, ten Wolde PR (2007) An allosteric model of circadian KaiC phosphorylation. *Proc Natl Acad Sci USA* 104:7420–7425
34. Reinitz J, Mjolsness E, Sharp DH (1995) Model for cooperative control of positional information in *Drosophila* by bicoid and maternal hunchback. *Exp J Zool* 271:47–56
35. von Dassow G, Meir E, Munro EM, Odell GM (2000) The segment polarity network is a robust developmental module. *Nature* 406:188–192
36. Kellershohn N, Laurent M (2001) Prion diseases: dynamics of the infection and properties of the bistable transition. *Biophys J* 81:2517–2529
37. Ferrell JEJ, Machleder EM (1998) The biochemical basis of an all-or-none cell fate switch in *Xenopus* oocytes. *Science* 280:895–898
38. Huang C-YF, Ferrell JEJ (1996) Ultrasensitivity in the mitogen-activated protein kinase cascade. *Proc Natl Acad Sci USA* 93:10078–10083
39. Laurent M, Kellershohn N (1999) Multistability: a major means of differentiation and evolution in biochemical systems. *Trends Biochem Sci* 24:418–422
40. Samoilov MS, Arkin AP (2006) Deviant effects in molecular reaction pathways. *Nature Biotech* 24:1235–1240
41. van Kampen NG (1992) *Stochastic Processes in Physics and Chemistry*. Elsevier, Amsterdam
42. Haseltine EL, Rawlings JB (2005) On the origins of approximations for stochastic chemical kinetics. *Chem J Phys* 123:164115
43. Gillespie DT (1992) A rigorous derivation of the chemical master equation. *Physica A* 188:404–425
44. Rohwer JM, Postma PW, Kholodenko BN, Westerhoff HV (1998) Implications of macromolecular crowding for signal transduction and metabolite channeling. *Proc Natl Acad Sci USA* 95:10547–10552
45. Andrews SS, Bray D (2004) Stochastic simulation of chemical reactions with spatial resolution and single molecule detail. *Phys Biol* 1:137–151
46. Munsky B, Khammash M (2006) The finite state projection algorithm for the solution of the chemical master equation. *Chem J Phys* 124:044104
47. Peles S, Munsky B, Khammash M (2006) Reduction and solution of the chemical master equation using time scale separation and finite state projection. *Chem J Phys* 125:204104
48. Kuwahara H, Myers CJ, Samoilov MS, Barker NA, Arkin AP (2006) Automated abstraction methodology for genetic regulatory networks. *Trans Comput Syst Biol* 6:150–175
49. Hegland M, Burden C, Santoso L, MacNamara S, Booth H (2007) A solver for the stochastic master equation applied to gene regulatory networks. *Comp J Appl Math* 205:708–724
50. Nedeia SV, Jansen APJ, Luskien JJ, Hilbers PAJ (2003) Infinitely fast diffusion in single-file systems. *Phys Rev E* 67:046707
51. Chatterjee A, Vlachos DG (2006) Multiscale spatial Monte Carlo simulations: Multigridding, computational singular per-



- turbation, and hierarchical stochastic closures. *Chem J Phys* 124:064110
52. Ambjörnsson T, Banik SK, Lomholt MA, Metzler R (2007) Master equation approach to DNA breathing in heteropolymer DNA. *Phys Rev E* 75:021908
  53. Altan-Bonnet G, Libchaber A, Krichevsky O (2003) Bubble dynamics in double-stranded DNA. *Phys Rev Lett* 90:138101
  54. Lattanzi G, Maritan A (2001) Master equation approach to molecular motors. *Phys Rev E* 64:061905
  55. Wang H-Y, Elston T, Mogilner A, Oster G (1998) Force generation in RNA polymerase. *Biophys J* 74:1186–1202
  56. Peskin CS, Odell GM, Oster GF (1993) Cellular motions and thermal fluctuations: the Brownian ratchet. *Biophys J* 65:316–324
  57. Paulsson J, Ehrenberg M (2000) Random signal fluctuations can reduce random fluctuations in regulated component of chemical regulatory networks. *Phys Rev Lett* 84:5447–5450
  58. Paulsson J, Berg OG, Ehrenberg M (2000) Stochastic focusing: fluctuation-enhanced sensitivity of intracellular regulation. *Proc Natl Acad Sci USA* 97:7148–7153
  59. Berg OG, Paulsson J, Ehrenberg M (2000) Fluctuations in repressor control: thermodynamic constraints on stochastic focusing. *Biophys J* 79:2944–2953
  60. Li H, Hou Z, Xin H (2005) Internal noise stochastic resonance for intracellular calcium oscillations in a cell system. *Phys Rev E* 71:061916
  61. Samoilov M, Pilyasunov S, Arkin AP (2005) Stochastic amplification and signaling in enzymatic futile cycles through noise-induced bistability with oscillations. *Proc Natl Acad Sci USA* 102:2310–2315
  62. Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Comp J Phys* 22:435–450
  63. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *Phys J Chem* 81:2340–2361
  64. Gibson MA, Bruck J (2000) Efficient exact stochastic simulation of chemical systems with many species and many channels. *Phys J Chem A* 104:1876–1889
  65. Lok L, Brent R (2005) Automatic generation of cellular reaction networks with Molecularizer 1.0. *Nature Biotech* 23:131–136
  66. Cao Y, Li H, Petzold L (2004) Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. *Chem J Phys* 121:4059–4067
  67. McCollum JM, Peterson GD, Cox CD, Simpson ML, Samatova NF (2006) The sorting direct method for stochastic simulation of biochemical systems with varying reaction execution behavior. *Comp Biol Chem* 30:39–49
  68. Pilyasunov S, Arkin AP (2007) Efficient stochastic sensitivity analysis of discrete event systems. *Comput J Phys* 221:724–738
  69. Bardwell L (2004) A walk-through of the yeast mating pheromone response pathway. *Peptides* 25:1465–1476
  70. Morton-Firth CJ, Bray D (1998) Predicting temporal fluctuations in an intracellular signalling pathway. *Theor J Biol* 192:117–128
  71. LeNovère N, Shimizu TS (2001) StochSim: modelling of stochastic biomolecular processes. *Bioinformatics* 17:575–576
  72. Lu T, Volfson D, Tsimring L, Hasty J (2004) Cellular growth and division in the Gillespie algorithm. *Syst Biol* 1:121–128
  73. McAdams H, Arkin A (1998) Simulation of prokaryotic genetic circuits. *Annu Rev Biophys Biomol Struct* 27:199–224
  74. Paulsson J (2004) Summing up the noise in gene networks. *Nature* 427:415–418
  75. McAdams HH, Arkin A (1997) Stochastic mechanisms in gene expression. *Proc Natl Acad Sci USA* 94:814–819
  76. Arkin A, Ross J, McAdams HH (1998) Stochastic kinetic analysis of developmental pathway bifurcation in phage lambda-infected *Escherichia coli* cells. *Genetics* 149:1633–1648
  77. Wolf DM, Vazirani VV, Arkin AP (2005) Diversity in times of adversity: probabilistic strategies in microbial survival games. *Theor J Biol* 234:227–253
  78. Fiering S, Whitelaw E, Martin DIK (2000) To be or not to be active: the stochastic nature of enhancer action. *BioEssays* 22:381–387
  79. Barkai N, Leibler S (2000) Biological rhythms: Circadian clocks limited by noise. *Nature* 403:267–268
  80. Thattai M, van Oudenaarden A (2004) Stochastic gene expression in fluctuating environments. *Genetics* 167:523–530
  81. Kussell E, Leibler S (2005) Phenotypic diversity, population growth, and information in fluctuating environments. *Science* 309:2075–2078
  82. Pedraza JM, van Oudenaarden A (2005) Noise propagation in gene networks. *Science* 307:1965–1969
  83. Swain PS, Elowitz MB, Siggia ED (2002) Intrinsic and extrinsic contributions to stochasticity in gene expression. *Proc Natl Acad Sci USA* 99:12795–12800
  84. Mettetal JT, Muzzey D, Pedraza JM, Ozbudak EM, van Oudenaarden A (2006) Predicting stochastic gene expression dynamics in single cells. *Proc Natl Acad Sci USA* 103:7304–7309
  85. Kierzek AM, Zaim J, Zielenkiewicz P (2001) The effect of transcription and translation initiation frequencies on the stochastic fluctuations in prokaryotic gene expression. *Biol J Chem* 276:8165–8172
  86. Peccoud J, Ycart B (1995) Markovian modeling of gene-product synthesis. *Theor Popul Biol* 48:222–234
  87. Rosenfeld N, Young JW, Alon U, Swain PS, Elowitz MB (2005) Gene regulation at the single-cell level. *Science* 307:1962–1965
  88. Kitano H (2004) Biological robustness. *Nature Rev Genet* 5:826–837
  89. Alon U, Surette MG, Barkai N, Leibler S (1999) Robustness in bacterial chemotaxis. *Nature* 397:168–171
  90. Barkai N, Leibler S (1997) Robustness in simple biochemical networks. *Nature* 387:913–917
  91. Stelling J, Sauer U, Szallasi Z, Doyle FJI, Doyle J (2004) Robustness of cellular functions. *Cell* 118:675–685
  92. Vilar JMG, Kueh HY, Barkai N, Leibler S (2002) Mechanisms of noise-resistance in genetic oscillators. *Proc Natl Acad Sci USA* 99:5988–5992
  93. Aldana M, Cluzel P (2003) A natural class of robust networks. *Proc Natl Acad Sci USA* 100:8710–8714
  94. Thattai M, van Oudenaarden A (2002) Attenuation of noise in ultrasensitive signaling cascades. *Biophys J* 82:2943–2950
  95. Yi T-M, Huang Y, Simon MI, Doyle J (2000) Robust perfect adaptation in bacterial chemotaxis through integral feedback control. *Proc Natl Acad Sci USA* 97:4649–4653
  96. Fraser HB, Hirsh AE, Giaever G, Kumm J, Eisen MB (2004) Noise minimization in eukaryotic gene expression. *PLoS Biol* 2:1–5
  97. Voigt CA, Wolf DM, Arkin AP (2005) The *Bacillus subtilis* sin operon: an evolvable network motif. *Genetics* 169:1187–1202

98. Cao Y, Gillespie DT, Petzold LR (2006) Efficient step size selection for the tau-leaping simulation method. *Chem J Phys* 124:044109
99. Gillespie DT, Petzold LR (2003) Improved leap-size selection for accelerated stochastic simulation. *Chem J Phys* 119:8229–8234
100. Gillespie DT (2001) Approximate accelerated stochastic simulation of chemically reacting systems. *Chem J Phys* 115:1716–1733
101. Cao Y, Gillespie DT, Petzold LR (2005) Avoiding negative populations in explicit Poisson tau-leaping. *Chem J Phys* 123:054104
102. Chatterjee A, Mayawala K, Edwards JS, Vlachos DG (2005) Time accelerated Monte Carlo simulations of biological networks using the binomial tau-leap method. *Bioinformatics* 21:2136–2137
103. Pettigrew MF, Resat H (2007) Multinomial tau-leaping method for stochastic kinetic simulations. *Chem J Phys* 126:084101
104. Zwanzig R (2001) A chemical Langevin equation with non-Gaussian noise. *Phys J Chem B* 105:6472–6473
105. Gillespie DT (1996) The multivariate Langevin and Fokker–Planck equations. *Am Phys J* 64:1246–1257
106. Gillespie DT (2000) The chemical Langevin equation. *Chem J Phys* 113:297–306
107. Gillespie DT (2002) The chemical Langevin and Fokker–Planck equations for the reversible isomerization reaction. *Phys J Chem A* 106:5063–5071
108. Wang H, Peskin CS, Elston TC (2003) A robust numerical algorithm for studying biomolecular transport processes. *Theor J Biol* 221:491–511
109. Xing J, Wang H, Oster G (2005) From continuum Fokker–Planck models to discrete kinetic models. *Biophys J* 89:1551–1563
110. Tao Y (2004) Intrinsic noise, gene regulation and steady-state statistics in a two-gene network. *Theor J Biol* 231:563–568
111. van der Mee CVM, Zweifel PF (1987) A Fokker–Planck equation for growing cell populations. *Math J Biol* 25:61–72
112. Sato K, Kaneko K (2006) On the distribution of state values of reproducing cells. *Phys Biol* 3:74–82
113. Hill NA, Häder D-P (1997) A biased random walk model for the trajectories of swimming micro-organisms. *Theor J Biol* 186:503–526
114. Schienbein M, Gruler H (1993) Langevin equation, Fokker–Planck equation and cell migration. *Bull Math Biol* 55:585–608
115. Xing J, Liao J-C, Oster G (2005) Making ATP. *Proc Natl Acad Sci USA* 102:16539–16546
116. Elston TC, Oster G (1997) Protein turbines I: the bacterial flagellar motor. *Biophys J* 73:703–721
117. Allen RJ, Frenkel D, ten Wolde PR (2006) Simulating rare events in equilibrium or nonequilibrium stochastic systems. *Chem J Phys* 124:024102
118. Rathinam M, Petzold LR, Cao Y, Gillespie DT (2003) Stiffness in stochastic chemically reacting systems: The implicit tau-leaping method. *Chem J Phys* 119:12784
119. Cao Y, Gillespie DT, Petzold LR (2005) The slow-scale stochastic simulation algorithm. *Chem J Phys* 122:014116
120. Cao Y, Gillespie DT, Petzold LR (2005) Accelerated stochastic simulation of the stiff enzyme-substrate reaction. *Chem J Phys* 123:144917
121. Rao CV, Arkin AP (2003) Stochastic chemical kinetics and the quasi-steady-state assumption: Application to the Gillespie algorithm. *Chem J Phys* 118:4999–5010
122. Adalsteinsson D, McMillen D, Elston TC (2004) Biochemical network stochastic simulator (BioNetS): software for stochastic modeling of biochemical networks. *Bioinformatics BMC* 5:24
123. Vasudeva K, Bhalla US (2004) Adaptive stochastic-deterministic chemical kinetic simulations. *Bioinformatics* 20:78–84
124. Baumeister W (2002) Electron tomography: towards visualizing the molecular organization of the cytoplasm. *Curr Opin Struct Biol* 12:679–684
125. Gierer A, Meinhardt H (1972) A theory of biological pattern formation. *Biol Cyber* 12:30–39
126. Maini PK, Painter KJ, Chau HNP (1997) Spatial pattern formation in chemical and biological systems. *Chem J Soc Faraday Trans* 93:3601–3610
127. Gurdon JB, Bourillot P-Y (2001) Morphogen gradient interpretation. *Nature* 413:797–803
128. Meinhardt H, de Boer PAJ (2001) Pattern formation in *Escherichia coli*: A model for the pole-to-pole oscillations of Min proteins and the localization of the division site. *Proc Natl Acad Sci USA* 98:14202–14207
129. Howard M, Rutenberg AD (2003) Pattern formation inside bacteria: fluctuations due to the low copy number of proteins. *Phys Rev Lett* 90:128102
130. Huang KC, Meir Y, Wingreen NS (2003) Dynamic structures in *Escherichia coli*: spontaneous formation of MinE rings and MinD polar zones. *Proc Natl Acad Sci USA* 100:12724–12728
131. Lutkenhaus J (2007) Assembly and dynamics of the bacterial MinCDE system and spatial regulation of the Z ring. *Ann Rev Biochem* 76:14.11–14.24
132. Fange D, Elf J (2006) Noise-induced Min phenotypes in *E. coli*. *PLoComp S Biol* 2:637–648
133. Kerr RA, Levine H, Sejnowski TJ, Rappel W-J (2006) Division accuracy in a stochastic model of Min oscillations in *Escherichia coli*. *Proc Natl Acad Sci USA* 103:347–352
134. Cytrynbaum E, Marshall BDL (2007) A multi-stranded polymer model explains MinDE dynamics in *E. coli* cell division. *Biophys J* 93:1134–1150
135. Bray D (1998) Signaling complexes: biophysical constraints on intracellular communication. *Annu Rev Biophys Biomol Struct* 27:59–75
136. Slepchenko BM, Schaff JC, Carson JH, Loew LM (2002) Computational cell biology: Spatiotemporal simulation of cellular events. *Annu Rev Biophys Biomol Struct* 31:423–441
137. Meyers J, Craig J, Odde DJ (2006) Potential for control of signaling pathways via cell size and shape. *Curr Biol* 16:1685–1693
138. Rao CV, Kirby JR, Arkin AP (2005) Phosphatase localization in bacterial chemotaxis: divergent mechanisms, convergent principles. *Phys Biol* 2:148–158
139. Agmon N, Edelstein AL (1997) Collective binding properties of receptor arrays. *Biophys J* 72:1582–1594
140. Lagerholm BC, Thompson NL (1998) Theory for ligand rebinding at cell membrane surfaces. *Biophys J* 74:1215–1228
141. Andrews SS (2005) Serial rebinding of ligands to clustered receptors as exemplified by bacterial chemotaxis. *Phys Biol* 2:111–122

142. Elf J, Ehrenberg M (2004) Spontaneous separation of bi-stable biochemical systems into spatial domains of opposite phases. *Syst Biol* 1:230–236
143. van Zon JS, ten Wolde PR (2005) Green's function reaction dynamics: A particle-based approach for simulating biochemical networks in time and space. *Chem J Phys* 123:234910
144. Stiles JR, Bartol TM (2001) Monte Carlo methods for simulating realistic synaptic microphysiology using MCell. In: De Schutter E (ed) *Computational Neuroscience: Realistic Modeling for Experimentalists*. Press CRC, Boca Raton
145. Dab D, Boon J-P, Li Y-X (1991) Lattice-gas automata for coupled reaction-diffusion equation. *Phys Rev Lett* 66:2535–2539
146. Ermentrout GB, Edelstein-Keshet L (1993) Cellular automata approaches to biological modeling. *Theor J Biol* 160:97–133
147. Duke TAJ, LeNovère N, Bray D (2001) Conformational spread in a ring of proteins: a stochastic approach to allostery. *Mol J Biol* 308:541–553
148. Goldman J, Andrews SS, Bray D (2004) Size and composition of membrane protein clusters predicted by Monte Carlo analysis. *Eur Biophys J* 33:506–512
149. Grima R, Schnell S (2006) A systematic investigation of the rate laws valid in intracellular environments. *Biophys Chem* 124:1–10
150. Turing AM (1990) The chemical basis of morphogenesis. *Bull Math Biol* 52:153–197
151. Cross MC, Hohenberg PC (1993) Pattern formation outside of equilibrium. *Rev Mod Phys* 65:851–1123
152. Slepchenko B, Schaff J, Macara I, Loew LM (2003) Quantitative cell biology with the Virtual Cell. *Cell TRENDS Biol* 13:570–576
153. Fink CC, Slepchenko B, Moraru II, Watras J, Schaff JC, Loew LM (2000) An image-based model of calcium waves in differentiated neuroblastoma cells. *Biophys J* 79:163–183
154. Fink CC, Slepchenko B, Moraru II, Schaff J, Watras J, Loew LM (1999) Morphological control of inositol-1,4,5-triphosphate-dependent signals. *Cell J Biol* 147:929–935
155. Hernjak N, Slepchenko B, Fernald K, Fink CC, Fortin D, Moraru II, Watras J, Loew LM (2005) Modeling and analysis of calcium signaling events leading to long-term depression in cerebellar Purkinje cells. *Biophys J* 89:3790–3806
156. Baras F, Malek-Mansour M (1996) Reaction-diffusion master equation: A comparison with microscopic simulations. *Phys Rev E* 54:6139–6148
157. Stundzia AB, Lumsden CJ (1996) Stochastic simulation of coupled reaction-diffusion processes. *Comput J Phys* 127:196–207
158. Nicolis G, Malek-Mansour M (1980) Systematic analysis of the multivariate master equation for a reaction-diffusion system. *Stat J Phys* 22:495–512
159. Kruse K, Elf J (2006) Kinetics in spatially extended systems. In: Szallasi Z, Stelling J, Periwal V (eds) *System Modeling in Cell Biology From Concepts to Nuts and Bolts*. Press MIT, Cambridge, pp 177–198
160. Hynes JT (1985) The theory of reactions in solution. In: Baer M (ed) *Theory of Chemical Reaction Dynamics*. Press CRC, Boca Raton, pp 171–234
161. Cohen B, Huppert D, Agmon N (2000) Non-exponential Smoluchowski dynamics in fast acid-base reaction. *Am J Chem Soc* 122:9838–9839
162. Noyes RM (1955) Kinetics of competitive processes when reactive fragments are produced in pairs. *Am J Chem Soc* 77:2042–2045
163. Pines E, Huppert D (1988) Geminate recombination in excited-state proton transfer reactions: Numerical solution of the Debye–Smoluchowski equation with backreaction and comparison with experimental results. *Chem J Phys* 88:5620–5630
164. Verkman AS (2002) Solute and macromolecule diffusion in cellular aqueous compartments. *Trends Biochem Sci* 27:27–33
165. Schnell S, Turner TE (2004) Reaction kinetics in intracellular environments with macromolecular crowding: simulations and rate laws. *Prog Biophys Mol Biol* 85:235–260
166. Fulton AB (1982) How crowded is the cytoplasm? *Cell* 30:345–347
167. Bernstein D (2005) Simulating mesoscopic reaction-diffusion systems using the Gillespie algorithm. *Phys Rev E* 71:041103
168. Ander M, Beltrao P, Di Ventura B, Ferkinghoff-Borg J, Foglierini M, Kaplan A, Lemerle C, Tomás-Oliveira I, Serrano L (2004) SmartCell, a framework to simulate cellular processes that combines stochastic approximation with diffusion and localisation: analysis of simple networks. *Syst Biol* 1: 129–138
169. Fricke T, Schnakenberg J (1990) Monte-Carlo simulation of an inhomogeneous reaction-diffusion system in the biophysics of receptor cells. *Z Phys B* 83:277–284
170. Isaacson SA, Peskin CS (2006) Incorporating diffusion in complex geometries into stochastic chemical kinetics simulations. *Sci SIAMJ Comput* 28:47–74
171. Hattne J, Fange D, Elf J (2005) Stochastic reaction-diffusion simulation with MesoRD. *Bioinformatics* 21:2923–2924
172. Frenkel D, Smit B (2002) *Understanding molecular simulation: from algorithms to applications*. Academic, San Diego
173. Berg HC (1993) *Random Walks in Biology*. Princeton Univ Press, Princeton
174. Gillespie DT (1996) The mathematics of Brownian motion and Johnson noise. *Am Phys J* 64:225–240
175. Rice SA (1985) *Diffusion Limited Reactions*. Elsevier, Amsterdam
176. Ermak DL, McCammon JA (1978) Brownian dynamics with hydrodynamic interactions. *Chem J Phys* 69:1352–1360
177. Northrup SH, Allison SA, McCammon JA (1984) Brownian dynamics simulation of diffusion-influenced bimolecular reactions. *Chem J Phys* 80:1517–1524
178. Northrup SH (1988) Diffusion-controlled ligand binding to multiple competing cell-bound receptors. *Phys J Chem* 92:5847–5850
179. Northrup SH, Erickson HP (1992) Kinetics of protein-protein association explained by Brownian dynamics computer simulation. *Proc Natl Acad Sci USA* 89:3338–3342
180. Edelstein AL, Agmon N (1993) Brownian dynamics simulations of reversible reactions in one dimension. *Chem J Phys* 99:5396–5404
181. Oh C, Kim H, Shin KJ (2002) Excited-state diffusion-influenced reversible association-dissociation reaction: Brownian dynamics simulation in three dimensions. *Chem J Phys* 117:3269–3277
182. Kim H, Yang M, Shin KJ (1999) Dynamic correlation effect in reversible diffusion-influenced reactions: Brownian dynamics simulation in three dimensions. *Chem J Phys* 111:1068–1075
183. Agmon N, Edelstein A (1995) Geometric and many-particle aspects of transmitter binding. *Biophys J* 68:815–825

184. Edelstein AL, Agmon N (1997) Brownian simulation of many-particle binding to a reversible receptor array. *Comput J Phys* 132:260–275
185. Agmon N, Szabo A (1990) Theory of reversible diffusion-influenced reactions. *Chem J Phys* 92:5270–5284
186. Kim H, Shin KJ (1999) Exact solution of the reversible diffusion-influenced reaction for an isolated pair in three dimensions. *Phys Rev Lett* 82:1578–1581
187. van Zon JS, ten Wolde PR (2005) Simulating biochemical networks at the particle level in time and space: Green's function reaction dynamics. *Phys Rev Lett* 94:128103
188. van Zon JS, Morelli MJ, Tanase-Nicola S, ten Wolde PR (2006) Diffusion of transcription factors can drastically enhance the noise in gene expression. *Biophys J* 91:4350–4367
189. Lipkow K (2006) Changing cellular location of CheZ predicted by molecular simulations. *Comp PLOS Biol* 2:301–310
190. Lipkow K, Andrews SS, Bray D (2004) Simulated diffusion of CheYp through the cytoplasm of *E. coli*. *J Bact* 187:45–53
191. Tournier AL, Fitzjohn PW, Bates PA (2006) Probability-based model of protein-protein interactions on biological timescales. *Algorithms Molec Biol* 1:25
192. Tolle DP, Le Novère N (2006) Particle-based stochastic simulation in systems biology. *Curr Bioinformatics* 1:1–6
193. Franks KM, Bartol TM, Sejnowski TJ (2002) A Monte Carlo model reveals independent signaling at central glutamatergic synapses. *Biophys J* 83:2333–2348
194. Coggan JS, Bartol TM, Esquenazi E, Stiles JR, Lamont S, Martone ME, Berg DK, Ellisman MH, Sejnowski TJ (2005) Evidence for ectopic neurotransmission at a neuronal synapse. *Science* 309:446–451
195. Koh X, Srinivasan B, Ching HS, Levchenko A (2006) A 3D Monte Carlo analysis of the role of dyadic space geometry in spark generation. *Biophys J* 90:1999–2014
196. Stiles JR, van Helden D, Thomas J, Bartol M, Salpeter EE, Salpeter MM (1996) Miniature endplate current rise times < 100 microseconds from improved dual recordings can be modeled with passive acetylcholine diffusion from a synaptic vesicle. *Proc Natl Acad Sci USA* 93:5747–5752
197. Stiles JR, Kovyazina IV, Salpeter EE, Salpeter MM (1999) The temperature sensitivity of miniature endplate currents is mostly governed by channel gating: Evidence from optimized recordings and Monte Carlo simulations. *Biophys J* 77:1177–1187
198. Bartol TMJ, Land BR, Salpeter EE, Salpeter MM (1991) Monte Carlo simulation of miniature endplate current generation in the vertebrate neuromuscular junction. *Biophys J* 59:1290–1307
199. Howard M, Rutenberg AD, de Vet S (2001) Dynamic compartmentalization of bacteria: accurate division in *E. coli*. *Phys Rev Lett* 87:278102
200. Kruse K (2002) A dynamic model for determining the middle of *Escherichia coli*. *Biophys J* 82:618–627
201. Wio HS (1996) Stochastic resonance in a spatially extended system. *Phys Rev E* 54:R3075–R3078
202. Hu Z, Gogol EP, Lutkenhaus J (2002) Dynamic assembly of MinD on phospholipid vesicles regulated by ATP and MinE. *Proc Natl Acad Sci USA* 99:6761–6766
203. Hu Z, Saez C, Lutkenhaus J (2003) Recruitment of MinC, an inhibitor of Z-ring formation, to the membrane in *Escherichia coli*: role of MinD and MinE. *J Bact* 185:196–203
204. Shih Y-L, Fu X, King GF, Le T, Rothfield L (2002) Division site placement in *E. coli*: mutations that prevent formation of the MinE ring lead to loss of the normal midcell arrest of growth of polar MinD membrane domains. *EMBO J* 21:3347–3357
205. Shih Y-L, Le T, Rothfield L (2003) Division site selection in *Escherichia coli* involves dynamic redistribution of Min proteins within coiled structures that extend between the two cell poles. *Proc Natl Acad Sci USA* 100:7865–7870
206. Shih Y-L, Kawagishi I, Rothfield L (2005) The MreB and Min cytoskeletal-like systems play independent roles in prokaryotic polar differentiation. *Mol Microbiol* 58:917–928
207. Suefuiji K, Valluzzi R, RayChaudhuri D (2002) Dynamic assembly of MinD into filament bundles modulated by ATP, phospholipids, and MinE. *Proc Natl Acad Sci USA* 99:16776–16781
208. Pavin N, Paljetak C, Krstic V (2006) Min-protein oscillations in *Escherichia coli* with spontaneous formation of two-stranded filaments in a three-dimensional stochastic reaction-diffusion model. *Phys Rev E* 73:021904
209. Adelman JL, Andrews SS (2004) Intracellular pattern formation: A spatial stochastic model of bacterial division site selection proteins MinProc CDE. Complex Systems Summer School Final Project Papers, Santa Fe Institute, Santa Fe
210. Drew DA, Osborn MJ, Rothfield LI (2005) A polymerization-depolymerization model that accurately generates the self-sustained oscillatory system involved in bacterial division site placement. *Proc Natl Acad Sci USA* 102:6114–6118
211. Andrews SS, Arkin AP (2007) A mechanical explanation for cytoskeletal rings and helices in bacteria. *Biophys J* 93:1872–1884
212. Lippincott-Schwartz J, Snapp E, Kenworthy A (2001) Studying protein dynamics in living cells. *Nat Rev Mol Cell Biol* 2:444–456

## Stochastic Noises, Observation, Identification and Realization with

GIORGIO PICCI  
Department of Information Engineering,  
University of Padua, Padua, Italy

### Article Outline

Glossary  
Introduction  
Stochastic Realization  
Wide-Sense Stochastic Realization  
Geometric Stochastic Realization  
Dynamical System Identification  
Future Directions  
Bibliography

### Glossary

**Wiener process** A *Wiener process* is a continuous time stochastic process  $w = \{w(t); t \in \mathbb{R}\}$  with continuous sample paths and stationary independent incre-



ments of zero mean and finite variance. The variance of the increments  $w(t) - w(s)$  must then have the form  $\sigma^2(t - s)$  and  $w(t) - w(s)$  must have a Gaussian distribution with mean zero (Levy's theorem [30]). The Wiener process is *normalized* (or *standard*) if  $\sigma^2 = 1$ . A  $p$ -dimensional Wiener process is a vector stochastic process having  $p$  components,  $w_k$ ;  $k = 1, \dots, p$ , which are independent Wiener processes. A *wide-sense*  $p$ -dimensional normalized Wiener process is a continuous time stochastic process  $w = \{w(t); t \in \mathbb{R}\}$  which is continuous in mean square and has stationary orthogonal increments; i. e.,

$$\mathbb{E}\{[w(t_1) - w(s_1)][w(t_2) - w(s_2)]'\} = I_p | [s_1, t_1] \cap [s_2, t_2] |$$

where  $\mathbb{E}$  denotes expectation,  $I_p$  is the  $p \times p$  identity matrix, the prime denotes transpose and  $|\cdot|$  denotes length (Lebesgue measure). This is a weaker concept than that of a Wiener process.

A Wiener process with  $w(0) = 0$  is sometimes called a *Brownian motion process*. Usually a Brownian motion is only defined on the half line  $\mathbb{R}_+$ .

Wiener processes and stochastic integrals with respect to a Wiener process, introduced by Wiener and Itô, are the basic building blocks of *stochastic calculus* [14,25]. Since stochastic integrals only depend on the increments of  $w$ , it is immaterial whether an arbitrary random vector is added to  $w$ . The equivalence class of (wide sense) Wiener processes differing from each other by the addition of an arbitrary (constant) random vector of finite variance (e. g.  $w(0)$ ), will still be called a Wiener process and denoted by the symbol  $dw$ .

**White noise** Continuous time *white noise* is the formal time derivative of a Wiener process. This derivative must be understood in a suitable distributional sense since it fails to exist as a limit in any of the standard senses of probability theory.

Much more standard is the notion of *discrete-time* (strict or wide-sense) normalized *white noise process*. It is a sequence of independent equally distributed (respectively orthogonal) random vectors  $w = \{w(t); t \in \mathbb{Z}\}$  with unit variance; i. e.,

$$\mathbb{E}\{w(t)w(s)'\} = I\delta_{ts} := \begin{cases} I & \text{if } s = t \\ 0 & \text{if } s \neq t \end{cases}.$$

This last condition alone characterizes wide-sense white noise. We shall always require that  $w$  should have zero mean.

### Stochastic differential and difference equations

Stochastic differential equations are equations that

define pathwise a continuous time stochastic process  $x$  by "local" evolution laws of the type

$$dx(t) = f(x)dt + G(x)dw(t),$$

where  $dw$  is a vector Wiener process. The equation should in reality be interpreted as an integral equation, the last term being a stochastic integral in the sense of Itô, see [14,25]. Under certain growth conditions on the coefficients the equation can be shown to have a unique solution which, in case  $f$  and  $G$  depend pointwise on  $x(t)$ , is a continuous Markov process, in fact a *diffusion process*.

In discrete time ( $t \in \mathbb{Z}$ ) a *stochastic difference equation* is an object of the type

$$x(t+1) = f(x(t))dt + G(x(t))w(t)$$

where now  $w$  is white noise. The solution of an equation of this type is a (discrete-time) Markov process.

**Stochastic dynamical systems** In continuous time, a (finite-dimensional) *stochastic system* is a pair  $(x, y)$  of vector, say  $n$  and  $m$  dimensional, stochastic processes satisfying equations of the type

$$\begin{aligned} dx(t) &= f(x(t))dt + G(x(t))dw(t) \\ dy(t) &= h(x(t))dt + J(x(t))dw(t). \end{aligned}$$

The process  $x$ , which is Markov, is called the *state* of the system while  $y$  is the *output* process. The differential notation in the second equation is merely to allow a possible additive white noise component in the output process so  $y$  is actually the integral of the variable which would physically be called the output of the system. If there is no additive white noise component ( $J \equiv 0$ ) this trick is unnecessary and the output variable is simply expressed as a memoryless function of the state,  $dy/dt = h(x(t))$ . A similar definition, with the obvious modifications, serves to introduce the concept of discrete-time stochastic systems.

The probabilistic essence of the concept of stochastic system is more generally captured in terms of *conditional independence* of the sigma-algebras induced by the underlying processes. From this it can be clearly seen that the state process at time  $t$  plays the role of *dynamic memory* of the system: the future and the past of the output and of the state sigma-algebras are conditionally independent given the present state  $x(t)$ . For a formal definition the reader is for example referred to p. 219 in [51]. We shall introduce these concepts in a wide-sense context later on.



## Introduction

In this article we discuss some general ideas which motivate the use of stochastic dynamical models in applied sciences. We discuss the inherent mathematical problem of stochastic model building, called *stochastic realization* and the statistical problem of *(dynamic) system identification*, i. e. procedures for estimating dynamic stochastic models starting from observed data. We shall in particular discuss linear stochastic state-space models with random inputs, their construction and the relevant identification techniques. We shall assume that the reader has some general background on stationary stochastic processes, statistics [12] and linear system theory as for example exposed in the textbook [24].

There are no physical laws telling us how to obtain stochastic descriptions of nature. The laws of mechanics, electromagnetism, fluid dynamics etc. are, by nature, essentially deterministic. A fundamental empirical observation is that stochastic models generally come about as *aggregate* descriptions of complicated large deterministic systems. Think for example of describing mathematically the outcome of a dice-throwing experiment. Doing this by the laws of physics involves an enormous number of complicated factors such as the mechanics of a cubic-shaped rigid body flying in the air after an initial impulse applied on a specific region of one of its six faces. To describe the trajectory (i. e. the motion of the barycenter, the body orientation and the angular momenta) one should take into account, besides gravity, lift phenomena due to portance, drag, added-mass effects etc., describing the interaction with the surrounding fluid. Then one should describe the discontinuous mechanics of the dice landing on an elastic surface etc. Modeling all of this by the known laws of physics is perhaps possible and we may in principle be able to set up a set of algebraic-differential equations allowing us to predict exactly (or “almost exactly”) which upper face of the dice will eventually show. This however, provided we knew an enormous number of geometrical and physical parameters of the system such as, for example, the dimensions of the dice, its initial position and orientation, its mass and moments of inertia, the location, direction and strength of the initial impulse, the density of the air surrounding the trajectory (depending on the temperature, humidity etc.), the possible air convection phenomena, the mechanical structure and geometry of the landing medium, etc. Indeed, this predictive model based on the (deterministic) laws of physics would require a very large number of equations and involve an unimaginable detailed prior knowledge of the parameters of the experiment. Trivially, the rudimentary stochastic model consist-

ing of a probability distribution on the six possible outcomes  $\{1, 2, \dots, 6\}$ , although not allowing an “exact”, i. e. “deterministic” prediction of the outcome, is incomparably simpler. One could venture to say that most stochastic models used in science and engineering come about, for the same reason, as simple aggregate descriptions of complicated deterministic systems. Thermodynamics, to name just one of the most conspicuous instances of this phenomenon, can be seen as stochastic aggregation of large ensembles of microscopic particles evolving in time according to the deterministic laws of mechanics [28,43].

Simplification is paid in terms of uncertainty and the predictions based on stochastic aggregate models are by nature uncertain. The modeling process hence requires us to quantify the uncertainty of aggregate modes. In a more precise sense, we should investigate their mathematical structure and describe how to construct them. This will be one of the main concerns of this article.

## A Brownian Particle in a Heat Bath

The process of constructing aggregate models can be elucidated by a famous example, the *Ford-Kac-Mazur model* [20] which is probably the simplest explicitly solvable example of a problem of this nature.

Consider the problem of describing the dynamics of a particle, called the *Brownian particle*, coupled to a “heat bath” consisting of a very large number of identical particles obeying the laws of classical Hamiltonian mechanics and moving under the influence of a quadratic potential. Mathematically, the ensemble can be described as a large number of coupled harmonic oscillators. The Brownian particle is the only particle of the ensemble which is assumed to be accessible to external observation. One assumes that the macroscopic observables of the system are the relative position  $q(t)$  and/or the momentum  $p(t)$  of the Brownian particle, at each instant of time.

It is shown in [20] that in the limit when the number of particles in the heat bath tends to infinity, the motion of the Brownian particle is governed (exactly!) by the stochastic differential equations,

$$\dot{q}(t) = p(t) \quad (1)$$

$$dp(t) = -ap(t)dt + kdW(t), \quad a > 0 \quad (2)$$

the second of which is the ubiquitous *Langevin equation* of statistical physics. This equation contains a dynamical *friction term*,  $(-ap(t))$ , and a forcing function term  $w(t)$  which is described mathematically as a Wiener process. Physically, these terms can be interpreted as the influence of the surrounding medium on the observed particle, the

friction accounting for the energy transferred from the particle to the heat bath and the forcing term  $w(t)$  representing the sum of infinitesimal “random shocks” inflicted to the particle by the surrounding medium.

The stochastic process describing the temporal evolution of the observables  $q(t)$  and  $p(t)$  is a Markov diffusion process (in fact, both  $p(t)$  and the two-dimensional vector process  $[q(t) \ p(t)]$  are Markov). The evolution of the macroscopic observables is hence *dissipative*, because of the friction term, and *irreversible*, since diffusion processes evolve in time by the action of a semigroup which cannot be inverted to become a group. This is in accordance with the basic postulates of (macroscopic) thermodynamics.

The Ford–Kac–Mazur example can be generalized [48,49,50,52]. Assume a microscopic system described by a Hamiltonian  $H$  on the phase space  $\Omega$ , a smooth manifold of very large dimension  $2N$ . The solution of the Hamilton equations determines the phase of the system, say configurations and momenta  $z(t) := [q(t) \ p(t)]'$  at each time  $t$ , uniquely in terms of the initial value  $z(0) \in \Omega$ . The correspondence defines a flow  $z(t) = \Phi(t)z(0)$  on the phase space which leaves invariant the total energy,  $H(z(t))$ , of the system. The microscopic evolution is in this sense, reversible and conservative. However, microscopic phenomena are *complex* in the sense that they involve an enormous number of interacting components and it is impossible to keep track of the dynamics of such a complex system. One must then resort to a statistical description.

A probability distribution on  $\Omega$  which is *invariant for the Hamiltonian flow*  $U(t)$  defines “thermal equilibrium”. It is known that in a finite-dimensional space any absolutely continuous  $U(t)$ -invariant probability measure admits a one parameter family of densities  $\rho(z)$  of the Maxwell–Boltzmann type, equal to a normalization constant times  $\exp[-\frac{1}{2\beta}H(z)]$ ,  $\beta > 0$ , being interpreted as the “absolute temperature”. For a quadratic Hamiltonian in a linear phase space these distributions are Gaussian. The choice of an initial phase  $z(0)$  of a system in thermal equilibrium can then be interpreted as a random choice of an elementary event in the elementary outcome space  $\{\Omega, \rho\}$ .

It follows that in thermal equilibrium any observable; i.e. any measurable function  $h: \Omega \rightarrow \mathbb{R}$  on the phase space can be regarded as a random variable. In fact, once combined with the (measure-preserving) Hamiltonian flow  $z(t) = U(t)z(0)$ , any measurable observable defines a *stationary stochastic process*. In general, the time evolution of any finite, say  $m$ -dimensional, family of observables  $\{h_1, \dots, h_m\}$  can be identified with a vector-valued  $m$ -dimensional stationary process  $\{y(t)\}$  on  $\Omega$ , with

components

$$y_k(t, z(0)) := h_k(U(t)z(0)) \quad (3)$$

where  $z(0) \in \Omega$  is the elementary event and  $U(t)$  the measure preserving group of transformations on the underlying probability space.

The main point of this story is that the “randomization” of the phase space may, under certain conditions, lead to a description of certain observable processes  $\{y(t)\}$  of (3) by a stochastic dynamical model of a *much simpler structure than the microscopic deterministic Hamiltonian description*. These could for example be representations of  $\{y(t)\}$  of the type  $y(t) = h(x(t))$  where  $\{x(t)\}$  a finite dimensional, say  $\mathbb{R}^n$ -valued Markov process and  $h$  is a suitable function  $h: \mathbb{R}^n \rightarrow \mathbb{R}^m$ . This description, in case the Markovian representation has a smaller complexity (dimension) than that of the microscopic description, would precisely meet both the general physical plausibility for a thermodynamic description and the mathematical requirement of reduction of complexity.

*Remark 1* This view generalizes the way of conceiving stochastic aggregation in the physical literature, e.g [31,32], where it is generally required that the observables themselves should be Markov and satisfy a Langevin equation.

In general the requirement that the observables should themselves be Markov processes is seldom a reasonable one. For example, in the Brownian particle example the macroscopic observable  $q(t)$  by itself need *not* satisfy any stochastic differential equation. Formally,  $q(t)$  is just a memoryless function of a two-dimensional Markov process  $x(t) := [q(t) \ p(t)]'$  and it is instead the “thermodynamic state”  $x(t)$  which satisfies a stochastic differential equation.

## Stochastic Realization

*Stochastic realization* is the abstract version of the aggregation problem mentioned above. One studies the problem of representing an  $m$ -dimensional stationary process  $\{y(t)\}$  as a *memoryless function of a Markov process*. In this setting the physical phase space and the Hamiltonian group  $\{U_t\}$  are generalized to an abstract probability space and to a measure preserving one-parameter group of transformations mapping  $\Omega$  onto itself (the shift group of the process). In this article we shall assume that  $\{y(t)\}$  is smooth with continuous values but the problem area also concerns stationary finite state-processes, see [5,19,46,66], which have important applications in communications and biology. In the present context, the question is when

does a given stationary,  $m$ -dimensional, stochastic process  $\{y(t)\}$  admit representations of the type  $y(t) = h(x(t))$  where  $\{x(t)\}$  is a Markov diffusion process taking values on a finite-dimensional state space  $X$  and  $h$  is say a continuous function from  $X$  to  $\mathbb{R}^m$ .

Since a smooth Markov diffusion process  $\{x(t)\}$  is a solution of a stochastic differential equation (Chap. VI, Par. 3 in [14]), any such process must admit representations as the output of a stochastic dynamical system of the type

$$dx(t) = f(x(t))dt + G(x(t))dw(t) \quad (4)$$

$$y(t) = h(x(t)) \quad (5)$$

where  $\{w(t)\}$  is a vector  $m$  Wiener process. This stochastic dynamical system generalizes the Langevin equation representation. For a Gaussian mean square continuous stationary process, the representation simplifies to a linear model of the form

$$dx(t) = Ax(t)dt + Bdw(t) \quad (6)$$

$$y(t) = Cx(t) \quad (7)$$

where  $A, B, C$  are constant matrices of appropriate dimensions. There may be a need for introducing an additive noise term also in the second (output) equation when the process  $y$  has stationary increments, a situation more general than stationarity, see [37].

Representations as the output of a stochastic dynamical system are called *stochastic realizations* or *state-space realizations* of the process  $y$  and the Markov process  $x(t)$  is accordingly called the *state process* of  $y$ , or of the realization.

Signal descriptions by a finite dimensional stochastic realization can be transformed into other equivalent forms such as ARMA models etc., which are also widely used in the engineering and econometric literature. It should be stressed that, irrespective of which particular equivalent form of the stochastic model, virtually all sequential signal processing, prediction and control algorithms (Kalman filtering to name just the most popular) require availability of a finite-dimensional realization of one form or another. Hence signal representation by a stochastic system is indeed a crucial prerequisite for system analysis and design in engineering and applied sciences.

A substantial amount of literature on the stochastic realization problem has appeared in the last four decades. The stationary (and stationary increments) Gaussian case is now covered by a rather complete linear theory, and is surveyed e. g. in [35,37]. For a more up to date account, see the forthcoming book [41]. The nonlinear realization problem is still underdeveloped, see [45,51,62].

## Wide-Sense Stochastic Realization

We shall discuss only the wide-sense version of the realization problem, where random variables and processes are described in terms of first- and second-order moments. Since first- and second-order moments individuate a Gaussian distribution uniquely, the theory will in particular cover the Gaussian case. Hereafter, wide-sense stationarity will be simply referred to as stationarity. Also, a “wide-sense Markov process” will simply be called a “Markov process”.

The wide-sense realization problem may look like a very particular modeling problem to deal with, but it is completely solvable and is general enough to reveal quite explicitly some of the important features of the solution set of stochastic realizations, some of which are quite unexpected. It is also motivated by its wide applicability since most times in practice the only reasonable way to describe a priori random phenomena is by second-order moments, say by covariances or spectra.

We advise the reader that in many applied fields one may want to construct dynamical system models involving also exogenous input (or decision) variables. The stochastic realization problem of constructing linear state-space representations of a stationary process with inputs has been studied in [11,53,54,55]. In the last two references one may find also material related to the germane problem of *identification with inputs*.

For reasons of space limitations we shall not touch upon this subject. A pointer to the relevant identification literature will be given in the final section of the article.

We shall discuss in some detail only stochastic realization of *discrete-time* random processes. An analogous (and in fact somewhat simpler) treatment can be given for continuous time processes [35] but we choose discrete-time since this theory makes contact with system identification which we shall discuss later in the section entitled “*Dynamical System Identification*”.

Given a vector (say  $m$ -dimensional) wide-sense stationary purely nondeterministic (purely nondeterministic will be abbreviated to p.n.d. in the following; this property is called *linear regularity* in the Russian literature, see [57]) process  $y = \{y(t)\}$  with  $t \in \mathbb{Z}$  (the integers), one wants to find representations of  $y$  in terms of a finite dimensional Markov process. In other words one wants to find linear representations of the stationary process  $y$ , of the form

$$x(t+1) = Ax(t) + Bw(t) \quad (8a)$$

$$y(t) = Cx(t) + Dw(t) \quad (8b)$$

and procedures for constructing them. Here  $\{w(t)\}$  is a vector,  $p$ -dimensional normalized white noise process, i. e.  $\mathbb{E}\{w(t)w(s)'\} = I\delta(t-s)$ ,  $\mathbb{E}\{w(t)\} = 0$ ,  $\delta$  being the Kronecker delta function. Note that a white noise  $w$  could be seen as a (degenerate) kind of Markov process. The reason for having the term  $Dw(t)$  in (8b) is to avoid the presence of such degenerate components in the dynamic equation for the state process  $x$  which would artificially increase its dimension. With this convention, in the extreme circumstance where  $y = w$ , the state dimension (of a minimal realization) is zero.

The linear representation (8a) involves auxiliary variables, i. e. random quantities which are not given as a part of the original data, especially the  $n$ -dimensional state process  $x$  (a stationary Markov process) and the generating white noise  $w$ . The peculiar properties of these processes actually embody the desired system structure (8a). Constructing these auxiliary processes is an essential part of the realization problem.

We shall restrict ourselves to representations (8a) for which

- $(A, B, C)$  is a minimal triplet; i. e.  $(A, B)$  is a reachable pair and  $(A, C)$  is an observable pair, see e. g. [24] for these notions;
- $\begin{bmatrix} B \\ D \end{bmatrix}$  has independent columns.

These are classical conditions of nonredundancy of the model [24] and entail no loss of generality. From standard spectral representation theory, see e. g., Chap. 1 in [57], it follows that Eq. (8a) admits stationary solutions if and only if the rows of the  $n \times p$  matrix  $(e^{i\theta}I - A)^{-1}B$  are square integrable functions of  $\theta \in [-\pi, \pi]$ . This is equivalent to the absence of poles of the function  $z \rightarrow (zI - A)^{-1}B$  on the unit circle. Equivalently, the eigenvalues of  $A$  of modulus one, if any, must be “unreachable” for the pair  $(A, B)$ .

When eigenvalues on the unit circle are present, to guarantee stationarity they must be simple roots of the minimal polynomial of  $A$ . These eigenvalues, necessarily in even number say  $2k$ , give then rise to a sum of  $k$  uncorrelated sinusoidal oscillations with random amplitude, the so-called *purely deterministic* component of the process. This purely deterministic component of the stationary Markov process  $x$  obeys a fixed undriven (i. e. deterministic) linear difference equation of the type  $x(t+1) = A_0x(t)$ , which is the restriction of (8a) to the modulus one eigenspace of  $A$ . The initial conditions are random variables independent of the driving noise  $w$ . Since there is no stochastic forcing term in this restricted state equation, it follows that the presence of a purely deterministic component in the Markov process necessar-

ily implies that the pair  $(A, B)$  must be nonreachable. The first of the two conditions of nonredundancy listed above hence excludes the presence of a purely deterministic component in the state, and therefore also in the output process  $y$ .

From this discussion it is immediately seen that there are stationary solutions  $x$  of the difference Eq. (8a) which are p.n.d. Markov processes if and only if  $A$  does not have eigenvalues of modulus one; i. e.

$$|\lambda(A)| \neq 1. \quad (9)$$

When  $|\lambda(A)| < 1$ , the representation (8a) is called *forward* or *causal*. When  $|\lambda(A)| > 1$  the representation is called *backward* or *anticausal*. Traditionally the causality of a representation, i. e. the condition  $|\lambda(A)| < 1$  has been regarded as being equivalent to stationarity. In fact, stationarity permits a whole family of representations (8a) of the (same) process  $x$ , where the matrices  $A$  can have very different spectra. Formulas for computing a backward representation starting from a forward one or vice versa are given in [34,36]. The two representations (in particular the relative  $A$  matrices) are related by a particular family of *linear state feedback* transformations [47].

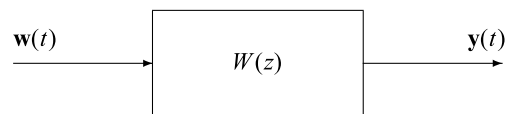
The *covariance function* of a zero-mean wide sense stationary process  $y$  is the  $m \times m$  matrix function  $k \mapsto \Lambda_k$ , defined by

$$\Lambda_k := \mathbb{E}\{y(t+k)y(t)'\} = \mathbb{E}\{y(k)y(0)'\} \quad k \in \mathbb{Z}.$$

This matrix function will be the initial data of our problem. Since  $y$  is p.n.d. the function  $\Lambda$  admits a Fourier transform called the *spectral density matrix* of  $y$ , given by

$$\Phi(z) = \sum_{t=-\infty}^{\infty} \Lambda_t z^{-t}$$

where  $z = e^{i\theta}$ ;  $\theta \in [-\pi, \pi]$ . It is easy to see that the spectral density matrix of a process admitting a state-space realization (8a) must be a rational function of  $z$ . This fact follows easily by eliminating the variable  $x$  in (8a) and thereby expressing  $y$  as the output of a linear filter as in Fig. 1 below whose transfer function  $W(z) = C(zI - A)^{-1}B + D$  is a rational function of  $z$ , and then using the classical Kintchine and Wiener formula for the spectrum of a filtered



Stochastic Noises, Observation, Identification and Realization with, Figure 1

Shaping filter representation of a stochastic process



stationary process [28,69] so that,

$$\Phi(z) = W(z)W(1/z)'. \quad (10)$$

Hence we have

**Proposition 1** *The transfer function  $W(z) = C(zI - A)^{-1}B + D$  of any state space representation (8a) of the stationary process  $y$ , is a spectral factor of  $\Phi$ , in the sense that it satisfies the spectral factorization equation (10).*

It is easy and instructive to check this directly. Supposing that one has a causal realization (8a), one can compute for  $k \geq 0$ ,

$$\begin{aligned} \Lambda_k &= \mathbb{E}[CA^k x(t) \\ &+ \sum_{s=0}^{k-1} CA^{k-1-s} Bw(t+s) + Dw(t+k)] \\ &\quad [Cx(t) + Dw(t)]' \end{aligned} \quad (11)$$

$$= CA^{k-1}\tilde{C}' + \frac{1}{2}\Lambda_0\delta(t), \quad (12)$$

where  $\tilde{C}' = \mathbb{E}x(t+1)y(t)'$  and  $\Lambda_0 = \mathbb{E}y(t)y(t)'$ . By identifying terms, these matrices can be expressed directly in terms of the realization parameters  $A, B, C, D$ , by

$$P = APA' + BB' \quad (13a)$$

$$\tilde{C}' = APC' + BD' \quad (13b)$$

$$\Lambda_0 = CPC' + DD' \quad (13c)$$

the matrix  $P = \mathbb{E}x(t)x(t)'$  being the covariance matrix of the state process. Note that  $P$  obeys Eq. (13a) since by causality ( $|\lambda(A)| < 1$ )  $x(t)$  is a function of the infinite past  $\{w(s); s < t\}$  and hence is uncorrelated with  $w(t)$  so that the two terms in the right hand side of (8a) are also uncorrelated and their variances add. In fact, by reachability of the model,  $P$  is actually (symmetric and) positive definite.

Now, introduce the decomposition

$$\Phi(z) = \Phi_+(z) + \Phi_+(1/z)' \quad (14)$$

where  $\Phi_+(z) = \frac{1}{2}\Lambda_0 + \Lambda_1 z^{-1} + \Lambda_2 z^{-2} + \dots$  is analytic outside of the unit circle. Clearly  $\Phi_+(z)$  must be the Fourier transform of the “causal” tract of the covariance function  $\Lambda$ , computed in (12). Hence

$$\Phi_+(z) = C(zI - A)^{-1}\tilde{C}' + \frac{1}{2}\Lambda_0. \quad (15)$$

This shows that the spectrum of a process  $y$  described by a state-space model (8a), besides being a rational function

of  $z$ , is directly expressible in parametric form in terms of the parameters of a causal realization. This explicit computation of the spectrum seems to be due to R.E. Kalman and B.D.O. Anderson [4,21].

“Classical” stochastic realization theory, developed in the late 1960s [3,4,17], deals with the inverse problem of computing the parameters  $A, B, C, D$  of a state space realization starting from the spectrum (or, equivalently, of the covariance function) of a stationary process  $y$ . Of course one assumes here that  $y$  is p.n.d. and has a spectral density  $\Phi$  which is a rational function of  $z = e^{j\theta}$ .

This inverse problem is just a parametric version of the spectral factorization problem: Given a rational spectrum  $\Phi(z)$  one looks for rational spectral factors parametrized as  $W(z) = C(zI - A)^{-1}B + D$ , solutions of (10). Since rational factors of a given spectral density matrix are in general infinitely many, one must preliminarily single out the interesting solutions by imposing certain restrictions. One very reasonable restriction which is imposed on the solutions of (10) is that the spectral factors should be of *minimal degree*. The degree of a proper rational matrix function is by definition the dimension  $n$  of the state of a minimal realization. Minimal degree (for short, *minimal*) spectral factors must have a degree which is exactly one half of the degree of the spectral density matrix  $\Phi(z)$ .

The solution of the inverse problem is described in the following proposition.

**Proposition 2** *Assume the spectrum is given in the parametric form (15) where  $|\lambda(A)| < 1$  and  $(A, C, \tilde{C}')$  is a minimal triplet of dimension  $n$ . Then the minimal degree spectral factors of  $\Phi(z)$ , analytic in  $\{z > 1\}$ , are in one-to-one correspondence with the symmetric  $n \times n$  matrices  $P$  solving the Linear Matrix Inequality (LMI)*

$$M(P) := \begin{bmatrix} P - APA' & \tilde{C}' - APC' \\ \tilde{C} - CPA' & \Lambda_0 - CPC' \end{bmatrix} \geq 0 \quad (16)$$

where the symbol  $\geq$  means positive semidefinite. The correspondence is to be understood in the following sense:

Corresponding to each solution  $P = P'$  of (16), consider the full column rank matrix factor  $\begin{bmatrix} B \\ D \end{bmatrix}$  of  $M(P)$ ,

$$M(P) = \begin{bmatrix} B \\ D \end{bmatrix} [B'D'] \quad (17)$$

(this factor is unique modulo right multiplication by orthogonal matrices) and form the rational matrix

$$W(z) := C(zI - A)^{-1}B + D. \quad (18)$$

Then (18) is a minimal realization of a minimal analytic spectral factor of  $\Phi(z)$  and all minimal analytic factors can be obtained in this way.



Note that any matrix  $P$  solving the LMI satisfies the system (12) and hence can be interpreted as the state covariance of a causal realization of  $y$ . It follows that any  $P$  solving the LMI is necessarily positive definite.

In fact it has been shown [17] that the set of symmetric solutions to the LMI (16)

$$\mathcal{P} := \{P = P'; M(P) \geq 0\}$$

is a closed, bounded and convex set with maximal and minimal elements  $P_-$ ,  $P_+ \in \mathcal{P}$  which satisfy

$$P_- \leq P \leq P_+ \quad \text{for all } P \in \mathcal{P}$$

where  $P_1 \leq P_2$  means that  $P_2 - P_1 \geq 0$  (is positive semi-definite).

The existence of solutions to the LMI is equivalent to the property of  $\Phi_+(z)$  defined in (14) of being a *positive-real matrix function*, namely that  $\Phi_+$  should be analytic in  $\{|z| > 1\}$  and that  $\Re \Phi_+(e^{j\theta}) \geq 0$ . The second condition is automatically satisfied when  $\Phi$  admits spectral factors since

$$2\Re \Phi_+(e^{j\theta}) = W(e^{j\theta})W(e^{j\theta})^* \geq 0$$

(the star meaning complex conjugate transpose). This property can be characterized in general, for not necessarily stable representations, by the so-called *positive-real lemma* discovered in the 1960s by Kalman, Yakubovich, and Popov [22,56,71] in the context of dissipative systems and stability theory. The first application of the Kalman–Yakubovich–Popov theory to factorization of spectral density functions and to stochastic realization (called *stationary covariance generation*) is due to B.D.O. Anderson [3,4]. Far reaching generalizations to not necessarily stable systems have been pursued by J.C. Willems in his widely quoted paper [70]. The terminology “positive real” derives from network synthesis.

Under certain conditions on the zeros of the spectrum, discussed e.g. in [18], the matrix  $\Delta(P) := \Lambda_0 - CPC'$  is nonsingular (and hence positive definite) for all  $P \in \mathcal{P}$ . In this case (16) is equivalent to the nonnegativity of the Schur complement of  $\Delta(P)$  in the matrix  $M(P)$ , which can be written

$$P - APA' - (\tilde{C}' - APC')\Delta(P)^{-1}(\tilde{C}' - APC')' \leq 0. \quad (19)$$

This is the *Algebraic Riccati Inequality* (ARI) of spectral factorization, first analyzed in the work of [3,17]. The solutions of the ARI with equality sign; i.e. the solutions of the Algebraic Riccati Equation (ARE)

$$P - APA' - (\tilde{C}' - APC')\Delta(P)^{-1}(\tilde{C}' - APC')' = 0 \quad (20)$$

make  $M(P)$  of minimal rank  $m = \dim y$  and hence correspond to *square spectral factors* (18). These spectral factors all have the same denominator matrix  $A$  and hence only differ by the location of their zeros. In particular one may show that  $P_-$ ,  $P_+$  solve the ARE and the two spectral factors  $W_-$  and  $W_+$  corresponding to  $P_-$ ,  $P_+$  (are square and) have all of their zeros inside and respectively outside of the unit circle. In other words,  $W_-$  is the *minimum phase* and  $W_+$  the *maximum phase* spectral factor. These spectral factors are essentially unique. All other square spectral factors are obtained (in rough terms) by “flipping” some of the zeros of  $W_-$  to their reciprocals outside of the unit circle. The maximum phase factor,  $W_+$ , is obtained by flipping *all* of the zeros of  $W_-$  to their reciprocals. More information on the zero structure of the minimal spectral factors can be found in [39].

In the next section we shall analyze in more detail the correspondence between spectral factors and stochastic realizations. As we shall see, for square spectral factors this correspondence is one-to-one and hence this particular class of stochastic realizations, for which  $\dim w = \dim y = m$  can be classified according to their zero location.

The LMI plays an important role in many areas of system theory such as stability theory, dissipative systems and is central in  $H^\infty$  control and estimation theory. It seems to be much less appreciated in the scientific community that it plays a very basic role in modeling of stationary random signals as well. As we shall see in the next section certain solutions of the LMI (or of the ARI) have special probabilistic properties and are related to Kalman-filter or “innovations-type” realizations.

## Geometric Stochastic Realization

The classical stationary covariance realization theory of the previous section is purely distributional as it says nothing about representation of random quantities in a truly probabilistic sense (i.e. how to generate the random variables or the sample paths of a given process, not just its covariance function). This was implicitly pointed out by Kalman already in [23]. In the last decades a *geometric coordinate-free* approach to stochastic modeling has been put forward in a series of papers by Lindquist, Picci, Ruckebusch et al. [33,34,35,58,59,60] which aims at the representation of random processes in this more specific sense. This motivation is also present in the early papers by Akaike [1,2].

A main point of the geometric approach is to provide procedures for the construction of the random quantities defining a stochastic state-space realization. In particular

the *state space* of a realization is defined in terms of the conditional independence relation between past and future of the signals involved. This relation is intrinsically *coordinate-free* and in the present setting involves only linear subspaces of a given ambient Hilbert space of random variables, typically made of linear functionals of the variables of the process  $y$  to be modeled (but in some situations other random data may be used to construct the model).

### Constructing the State Space of a Stationary Process

The theme of this section will be a review of geometric realization theory for the stochastic process  $y$ . The geometric theory centers on the idea of *Markovian Splitting Subspace* for the process  $y$ . This concept is the probabilistic analog of the deterministic notion of state space of a dynamical system and captures at an abstract level the property of “dynamic memory” that the state variables have in deterministic dynamical system theory. Once a stochastic state space is given the construction of the auxiliary random quantities which enter in a state space model and in particular the state process is fairly obvious. The state vector  $x(t)$  of a particular realization can be regarded just as a particular basis for the state space, hence once a state-space is constructed, finding state equations is just a matter of choosing a basis and computing coordinates.

**Notation 1** In what follows the symbols  $\vee$ ,  $+$  and  $\oplus$  will denote vector sum, direct vector sum and orthogonal vector sum of subspaces, the symbol  $\mathbf{x}^\perp$  will denote the orthogonal complement of a (closed) subspace  $\mathbf{X}$  of a Hilbert space with respect to some predefined ambient space. Given a collection  $\{X_\alpha \mid \alpha \in A\}$  of subsets of a Hilbert space  $\mathbf{H}$  we shall denote by  $\overline{\text{span}}\{X_\alpha \mid \alpha \in A\}$  the closure in  $\mathbf{H}$  of the linear (real) vector space generated by the collection. The orthogonal projection onto the subspace  $\mathbf{X}$  will be denoted by the symbol  $\mathbb{E}(\cdot \mid \mathbf{X})$  or by the shorthand  $\mathbb{E}^\mathbf{X}$ . The notation  $\mathbb{E}(z \mid \mathbf{X})$  will be used also when  $z$  is vector-valued. The symbol will then denote the vector with components  $E(z_k \mid \mathbf{X})$ ,  $k = 1, \dots$ . For vector quantities,  $|v|$  will denote Euclidean length (or absolute value in the scalar case).

Let  $y$  be a zero mean stationary vector process with finite second-order moments defined on some underlying probability space  $\{\Omega, \mathcal{A}, \mu\}$  and let  $L^2\{\Omega, \mathcal{A}, \mu\}$  denote the Hilbert space of second-order random variables defined on  $\Omega$ , endowed with the inner product  $\langle \xi, \eta \rangle = \mathbb{E}\{\xi\eta\}$ . Let  $\mathbf{H}(y)$  be the (closed) linear subspace of  $L^2\{\Omega, \mathcal{A}, \mu\}$ , linearly generated by the variables of the process; i. e.

$$\mathbf{H}(y) := \overline{\text{span}}\{y_k(t); k = 1, 2, \dots, m; t \in \mathbb{Z}\}.$$

If  $y$  is generated by a linear model of the type (8a) this Hilbert space will in general be a proper subspace of the

space  $\mathbf{H}(w)$ , generated by the input white noise of the model (8a). All random variables of the stochastic system (8a) belong to  $\mathbf{H}(w)$  and for this reason  $\mathbf{H}(w)$  is called the *ambient space* of the model. By stationarity of  $w$ ,  $\mathbf{H}(w)$  comes naturally equipped with a unitary operator  $U$ , called the *shift* of  $w$ , such that  $Uw_i(t) = w_i(t+1)$  for  $i = 1, 2, \dots, p$  and all  $t \in \mathbb{Z}$  (stationarity). Note that the family  $\{U^t; t \in \mathbb{Z}\}$  forms a one parameter group of unitary operators on the ambient space; in particular,  $U^{-t} = (U^*)^t$ . The pair  $(\mathbf{H}(w), U)$  is also called a *stationary Hilbert space*. Clearly the processes  $x$  and  $y$  will also be stationary with respect to  $U$ .

The *past* and *future* subspaces (at time zero) of a stationary process are

$$\mathbf{H}^-(y) := \overline{\text{span}}\{y_k(t); k = 1, 2, \dots, m; t < 0\}$$

$$\mathbf{H}^+(y) := \overline{\text{span}}\{y_k(t); k = 1, 2, \dots, m; t \geq 0\}.$$

Because of stationarity everything propagates in time by the action of the unitary group, e.g. the past and future at time  $t$  are obtained as  $\mathbf{H}_t^-(y) = U^t \mathbf{H}^-(y)$  and  $\mathbf{H}_t^+(y) = U^t \mathbf{H}^+(y)$ . For this reason the geometric definitions and constructions to follow, although valid for an arbitrary time instant will usually be referred to the time instant  $t = 0$ .

Let  $\mathbf{X}$  be a subspace of some large stationary Hilbert space  $\mathbf{H}$  of wide-sense random variables containing  $\mathbf{H}(y)$ . Define

$$\mathbf{X}_t := U^t \mathbf{X}, \quad \mathbf{X}_t^- := \vee_{s \leq t} \mathbf{X}_s, \quad \mathbf{X}_t^+ := \vee_{s \geq t} \mathbf{X}_s.$$

**Definition 1** A *Markovian Splitting Subspace*  $\mathbf{X}$  for the process  $y$  is a subspace of  $\mathbf{H}$  making the vector sums  $\mathbf{H}(y)^- \vee \mathbf{X}^-$  and  $\mathbf{H}(y)^+ \vee \mathbf{X}^+$  conditionally orthogonal (i. e. conditionally uncorrelated) given  $\mathbf{X}$ , denoted,

$$\mathbf{H}(y)^- \vee \mathbf{X}^- \perp \mathbf{H}(y)^+ \vee \mathbf{X}^+ \mid \mathbf{X}. \quad (21)$$

The conditional orthogonality condition (21) can be equivalently written as

$$\begin{aligned} \mathbb{E}[\mathbf{H}(y)^+ \vee \mathbf{X}^+ \mid \mathbf{H}(y)^- \vee \mathbf{X}^-] = \\ \mathbb{E}[\mathbf{H}(y)^+ \vee \mathbf{X}^+ \mid \mathbf{X}] \end{aligned} \quad (22)$$

$$\begin{aligned} \mathbb{E}[\mathbf{H}(y)^- \vee \mathbf{X}^- \mid \mathbf{H}(y)^+ \vee \mathbf{X}^+] = \\ \mathbb{E}[\mathbf{H}(y)^- \vee \mathbf{X}^- \mid \mathbf{X}] \end{aligned} \quad (23)$$

which formalize the intuitive meaning of the splitting subspace  $\mathbf{X}$  as a dynamic memory of the past (future) for the purpose of predicting the joint future (past). It follows in particular that  $\mathbf{X}$  is Markovian (i. e.  $\mathbf{X}^- \perp \mathbf{X}^+ \mid \mathbf{X}$ ) and

any basis vector  $x := [x_1, x_2, \dots, x_n]'$  in a (finite-dimensional) Markovian splitting subspace  $\mathbf{X}$  generates a stationary Markov process  $x(t) := U^t x$ ,  $t \in \mathbb{Z}$  which, as we shall see in a moment, serves as a *state* for the process  $y$ .

The subspace  $\mathbf{X}$  is called *proper*, or *p.n.d.* if

$$\cap_t \mathbf{Y}_t^- \vee \mathbf{X}_t^- = \{0\}, \text{ and } \cap_t \mathbf{H}(y)_t^+ \vee \mathbf{X}_t^+ = \{0\}.$$

Obviously for the existence of proper splitting subspaces  $y$  must also be purely nondeterministic [57]. Properness is, by the Wold decomposition theorem, equivalent to the existence of two vector white noise processes  $w$  and  $\bar{w}$  such that,

$$\mathbf{H}(y)^- \vee \mathbf{X}^- = \mathbf{H}^-(w), \quad \mathbf{H}(y)^+ \vee \mathbf{X}^+ = \mathbf{H}^+(\bar{w}).$$

If  $\mathbf{X}$  is proper, every Markov process  $x(t) := U^t x$  is purely nondeterministic.

The subspaces

$$\mathbf{S} := \mathbf{H}(y)^- \vee \mathbf{X}^- \text{ and } \bar{\mathbf{S}} := \mathbf{Y}^+ \vee \mathbf{X}^+ \quad (24)$$

associated to a Markovian Splitting subspace  $\mathbf{X}$ , play an important role in the geometric theory of stochastic systems. They are called the *scattering pair* of  $\mathbf{X}$  as they can be seen to form an incoming–outgoing pair in the sense of Lax–Phillips Scattering Theory [29].

**Definition 2** Given a stationary Hilbert space  $(\mathbf{H}, U)$  containing  $\mathbf{H}(y)$ , a *scattering pair* for the process  $y$  is a pair of subspaces  $(\mathbf{S}, \bar{\mathbf{S}})$  satisfying the following conditions for a stationary process,

1.  $U^* \mathbf{S} \subset \mathbf{S}$  and  $U \bar{\mathbf{S}} \subset \bar{\mathbf{S}}$ , i.e.  $\mathbf{S}$  and  $\bar{\mathbf{S}}$  are invariant for the left and right shift semigroups (this means that  $\mathbf{S}_t$  is increasing and  $\bar{\mathbf{S}}_t$  is decreasing with time).
2.  $\mathbf{S} \vee \bar{\mathbf{S}} = \mathbf{H}$
3.  $\mathbf{S} \supset \mathbf{H}(y)^-$  and  $\bar{\mathbf{S}} \supset \mathbf{H}(y)^+$
4.  $\mathbf{S}^\perp \subset \bar{\mathbf{S}}$  or, equivalently,  $\bar{\mathbf{S}}^\perp \subset \mathbf{S}$

The following representation Theorem provides the link between Markovian splitting subspaces and scattering pairs.

**Theorem 1 ([35])** *The intersection*

$$\mathbf{X} = \mathbf{S} \cap \bar{\mathbf{S}} \quad (25)$$

of any scattering pair of subspaces of  $\mathbf{H}$  is a Markovian splitting subspace. Conversely every Markovian splitting subspace can be represented as the intersection of a scattering pair. The correspondence  $\mathbf{X} \leftrightarrow (\mathbf{S}, \bar{\mathbf{S}})$  is one-to-one, the scattering pair corresponding to  $\mathbf{X}$  being given by

$$\mathbf{S} = \mathbf{H}(y)^- \vee \mathbf{X}^- \quad \bar{\mathbf{S}} = \mathbf{H}(y)^+ \vee \mathbf{X}^+. \quad (26)$$

The process of forming scattering pairs associated to  $\mathbf{X}$  should be thought of as an “extension” of the past and future spaces of  $y$ . The rationale for this extension is that scattering pairs have an extremely simple splitting geometry due to the fact that

$$\mathbf{S} \perp \bar{\mathbf{S}} \mid \mathbf{S} \cap \bar{\mathbf{S}} \quad (27)$$

equivalent to

$$\mathbf{S} \vee \bar{\mathbf{S}} = \bar{\mathbf{S}}^\perp \oplus (\mathbf{S} \cap \bar{\mathbf{S}}) \oplus \mathbf{S}^\perp \quad (28)$$

which is called *perpendicular intersection*. It is easy to show that Property (4) in the definition of a scattering pair is actually equivalent to perpendicular intersection. This property of conditional orthogonality given the intersection can also be seen as a natural generalization of the Markov property. Indeed for a Markovian family of subspaces  $\mathbf{X} = \mathbf{X}^- \cap \mathbf{X}^+$  and  $\mathbf{S} = \mathbf{X}^-$ ,  $\bar{\mathbf{S}} = \mathbf{X}^+$  intersect perpendicularly with intersection  $\mathbf{X}$ .

Note that  $\mathbf{A} \perp \mathbf{B} \mid \mathbf{X} \Rightarrow \mathbf{A} \cap \mathbf{B} \subset \mathbf{X}$  but the inclusion of the intersection in the splitting subspace  $\mathbf{X}$  is only *proper* in general. For perpendicularly intersecting subspaces, the intersection is actually the *unique minimal subspace* making them conditionally orthogonal.

Theorem 1 is the fundamental device for the construction and classification of Markovian splitting subspaces.

Denote by  $\mathbf{W}_t, \bar{\mathbf{W}}_t$  the (wandering) subspaces spanned by the components, at time  $t$ , of the generating noises  $w(t)$  and  $\bar{w}(t)$ , of the scattering pair of  $\mathbf{X}$ . Since

$$\mathbf{S}_{t+1} = \mathbf{S}_t \oplus \mathbf{W}_t, \quad (29)$$

we can write,

$$\mathbf{X}_{t+1} = \mathbf{S}_{t+1} \cap \bar{\mathbf{S}}_{t+1} \subset (\mathbf{S}_t \cap \bar{\mathbf{S}}_{t+1}) \oplus (\mathbf{W}_t \cap \bar{\mathbf{S}}_{t+1}) \quad (30)$$

Since  $\bar{\mathbf{S}}_t$  is decreasing in time, we have  $\mathbf{S}_t \cap \bar{\mathbf{S}}_{t+1} \subset \mathbf{X}_t$  and by projecting the shifted basis  $Ux(t) := x(t+1)$ , onto the last orthogonal direct sum above, the time evolution of any basis vector  $x(t) := [x_1(t), x_2(t), \dots, x_n(t)]'$  in  $\mathbf{X}_t$ , is described by a linear equation of the type

$$x(t+1) = Ax(t) + Bw(t).$$

It is also easy to see that, by the p.n.d. property,  $A$  must have all its eigenvalues strictly inside of the unit circle. Naturally, by decomposing instead  $\bar{\mathbf{S}}_{t-1} = \bar{\mathbf{S}}_t \oplus \bar{\mathbf{W}}_t$  one could have obtained a *backward difference equation* model for the Markov process  $x$ , driven by the backward generating noise  $\bar{w}$ .

To complete the representation, note that by definition of the past space,  $y(t) \in (\mathbf{S}_{t+1} \cap \bar{\mathbf{S}}_t)$ . Inserting the decomposition (29) and projecting  $y(t)$ , leads to a state-output

equation of the form

$$y(t) = Cx(t) + Dw(t).$$

Here one could also obtain a state-output equation driven by the backward noise  $\bar{w}$ , the same noise driving the backward state model obtained before.

As we have just seen, any basis in a Markovian splitting subspaces produces a stochastic realization of  $y$ . It is easy to reverse the implication. In fact the following fundamental characterization holds.

**Theorem 2** ([33,37]) *The state space  $\mathbf{X} = \text{span}\{x_1(0), x_2(0), \dots, x_n(0)\}$  of any stochastic realization (8a) is a Markovian splitting subspace for the process  $y$ .*

*Conversely, given a finite-dimensional Markovian splitting subspace  $\mathbf{X}$ , to any choice of basis  $x(0) = [x_1(0), x_2(0), \dots, x_n(0)]'$  in  $\mathbf{X}$  there corresponds a stochastic realization of  $y$  of the type (8a).*

Once a basis in  $\mathbf{X}$  is available, there are obvious formulas to compute the parameters  $(A, C, \bar{C})$  of the corresponding stochastic realization, namely

$$A = \mathbb{E}x(t+1)x(t)' P^{-1} \quad (31)$$

$$C = \mathbb{E}y(t)x(t)' P^{-1} \quad (32)$$

$$\bar{C} = \mathbb{E}y(t-1)x(t)' \quad (33)$$

where  $P = \mathbb{E}x(t)x(t)'$  is the state covariance matrix (i. e. the Gramian matrix of the basis). The matrices  $B$  and  $D$ , however, are related to the (unobservable) generating white noise  $w$  and require the solution of the LMI. This abstract procedure which permits us to compute the parameters of a stochastic realization once the state has been constructed, can be rendered quite concrete and forms the conceptual basis of subspace identification.

Stochastic realizations are called *internal* when  $\mathbf{H} = \mathbf{H}(y)$ , i. e. the state space is built from the Hilbert space made just of the linear statistics of the process  $y$ . For identification the only realizations of interest are the internal ones.

A central problem of geometric realization theory is to construct and to classify all minimal state spaces, i. e. the minimal Markovian splitting subspaces for the process  $y$ .

The obvious ordering of subspaces of  $\mathbf{H}$  by inclusion, induces an ordering on the family of Markovian splitting subspaces. The notion of minimality is most naturally defined with respect to this ordering. Note that this definition is independent of assumptions of finite-dimensionality and applies also to infinite-dimensional Markovian splitting subspaces, i. e. to situations where comparing dimension would not make much sense.

**Definition 3** A Markovian splitting subspace is *minimal* if it does not contain (properly) other Markovian splitting subspaces.

The study of minimality forms an elegant chapter of stochastic system theory. There are several known geometric and algebraic characterizations of minimality of splitting subspaces and of the corresponding stochastic state-space realizations. Since, however, the discussion of this topic would take us too far from the main theme of the paper we shall refer the reader to the literature [35,37].

Contrary to the deterministic situation minimal Markovian splitting subspaces are *nonunique*. Two very important examples are the *forward and backward predictor spaces* (at time zero):

$$\mathbf{X}_- := \mathbf{E}^{\mathbf{H}^-} \mathbf{H}^+ \quad \mathbf{X}_+ := \mathbf{E}^{\mathbf{H}^+} \mathbf{H}^- \quad (34)$$

for which we have the following characterization [35].

**Proposition 3** *The subspaces  $\mathbf{X}_-$  and  $\mathbf{X}_+$  are minimal Markovian splitting subspaces contained in the past  $\mathbf{H}^-$ , and, respectively, in the future  $\mathbf{H}^+$ , of the process  $y$ .*

A basis in the forward predictor space  $\mathbf{X}_-$  originates a stationary state-space model in which the state variables are linear functionals of the past history of the process  $y$ , i. e.  $x(t) \in \mathbf{H}_t^-(y)$ . In other words the state coincides with its best estimate (the orthogonal projection),  $E[x(t) | \mathbf{H}_t^-(y)]$  given the past of  $y$ . It follows that the dynamical equations (8a) describe in this case a steady-state Kalman predictor and the input white noise  $w = w_-$  is the steady state *innovation process* of  $y$ .

Before closing this section we should remark that a scattering picture emerges also in some of the early papers on the “unitary dilation” approach to statistical mechanics, e. g. [31,32]. It is remarkable that the abstract scattering picture described in this section makes contact with this literature. The dilation approach, however, deals (in our language) with the reverse problem of describing the stationary shift group starting from (Markov) processes which are described by a Langevin-type equation. This is a curious and (in our view) somewhat unnatural point of view which apparently does not lead to capture the phenomenon of *nonuniqueness of the macroscopic description* of the observables. Whether this has any physical relevance is however not clear to us.

## Dynamical System Identification

System identification is the problem of describing an observed time series (e. g. a sequence of real numbers representing stock prices, temperatures in a room, sampled

audio or video signals, etc.) by a linear dynamic model, in particular by a state-space model of the type (8a). It is a widely accepted viewpoint that system identification should be regarded as a *statistical* problem [40,42]. This is so because the experimental data which one wants to model very often come from complicated, often unknown, nonlinear physical phenomena subject to unknown interactions with the environment, and it would in general be impossible to attempt a “physical” modelization from first principles. Consequently the models describing the data should be stochastic. The general goal is to get accurate models which are as simple as possible. The problem can then be approached from (at least) two conceptually different viewpoints.

### Identification by Parametric Optimization

This is the mainstream “optimization” approach, based on the principle of minimizing a suitable measure of discrepancy between the observed data and the data predicted by a probability law underlying a certain chosen model class. Well-known examples of distance functions are the *likelihood function*, or the average squared *prediction-error* of the observed data corresponding to a particular model. Except for rather trivial model classes, these distance functions depend nonlinearly on the model parameters and the minimization can only be done numerically. Hence the optimization approach leads to iterative algorithms in the parameter space, say in the space of minimal  $(A, B, C, D)$  matrix quadruples which parametrize a chosen model class. In spite of the fact that this has been almost the only accepted paradigm in system identification for many decades, [42,61], this approach has several drawbacks, including the need of unique parametrization of multivariable models, the fact that the cost function generally has complicated local minima which, for moderate or large dimension of the model are very difficult to detect, and the inherent insensitivity of the cost to variations of some parameters and the corresponding ill-posedness of the estimation problem.

It seems that these limitations are a consequence of the intrinsically “blind” philosophy which underlies setting the problem as a parameter optimization problem. Almost all problems of controller and/or estimator design could in principle be formulated just as parametric optimization after choosing a certain family of parametric structures for the controller or the estimator. Pushing this philosophy to the extreme, in principle one would not need the maximum principle, Kalman filtering,  $H^\infty$  theory, etc. (in fact one would not need system and control theory altogether), since everything could be reduced to a nonlinear program-

ming problem in a suitable space of controller or estimator parameters. It is dubious however, whether any real progress in the field of control and estimation could have occurred by following this type of paradigm.

### Identification by Stochastic Realization

This is commonly referred to as “subspace methods” identification or “subspace identification” *tout-court*. Subspace identification is essentially a sample version of stochastic realization. Under a reasonable assumption of second-order ergodicity of the process  $y$  which has produced the observed data, one can set up a “sample” counterpart of the abstract Hilbert space operations of stochastic realization theory. In this setting the geometric operations of stochastic realization can be translated into statistical operations on the observed data. One obtains a statistical theory of model building which is in a sense perfectly isomorphic to the abstract probabilistic realization theory. The main point is the idea of constructing first a (sample) *state space* for the process  $y$ , starting from certain vector spaces, called the *future* and *past* spaces associated with the observations. According to the recipe of (34) the state space is constructed by orthogonal projection of the future onto the past, as seen in the previous section. Successively, a well conditioned basis is chosen in the state space e.g. by principal components (canonical correlation) analysis. Once a “robust” basis; i.e. a state vector, is chosen, the parameters  $A, C, \bar{C}$  of the model are computed by formulas analogous to (31),(32),(33). The final step of finding the  $B$  and  $D$  matrices requires solving a Riccati equation.

By similarity with realization theory, one recognizes that the inherent nonlinearity of model identification has to do with the quadratic nature of the spectral factorization problem. As we have seen, spectral factorization for state-space models involves the solution of a Riccati equation (or more generally of a linear matrix inequality), a problem which has been the object of intensive theoretical and numerical studies in the past three decades. The nonlinearity of the stochastic system identification problem is hence of a well-known and well-understood kind and is much better dealt with by the explicit methods of Riccati solution developed in system theory rather than by nonspecific optimization algorithms. This has allowed a systematic introduction of very reliable and efficient numerical linear algebra tools to solve the problem. The “subspace” approach has been suggested more or less implicitly by several authors in the past [6,13,15,44] but is first clearly presented in [63]. Various extensions to cover identification of systems with inputs have appeared since this paper and there



is now a large literature on the subject. For references we refer the reader to the recently published book [27].

Since the optimization approach is well covered in the literature, we shall just analyze subspace identification below.

### The Hilbert Space of a Time Series

The crucial conceptual step in subspace identification is a proper identification of the Hilbert space of random data in which modeling takes place. We shall briefly illustrate this point, following [38].

We shall initially consider an idealized situation in which the observed data (time series)

$$\{y_0, y_1, \dots, y_t, \dots\} \quad y_t \in \mathbb{R}^m \quad (35)$$

is infinitely long.

We shall assume that the limit for  $N \rightarrow \infty$  and for any  $\tau \geq 0$ , of the time averages

$$\frac{1}{N+1} \sum_{t=0}^N y_{t+\tau} y_t' \quad (36)$$

exists. It can be shown that this limit is a function  $\tau \rightarrow \Lambda_\tau$  of *positive type*. In the continuous-time setting, functions admitting an “ergodic” limit of the sample correlation function (36), have been studied in depth by Wiener in his famous work on Generalized Harmonic Analysis. Although a systematic translation of the continuous-time results of Wiener into discrete-time seems not to be available in the literature, it is quite evident that a totally analogous set of results holds also for discrete-time signals. In particular it is rather easy to show, by adapting Wiener’s proof for continuous time, that the limits of the time averages (36) form a matrix function  $\Lambda$  of positive type, in other words a *bona-fide stationary covariance matrix*, see [67,68] which can then be identified with the “true” covariance function of an underlying stochastic process.

The assumption actually implies that

$$\frac{1}{N+1} \sum_{t=t_0}^{N+t_0} y_{t+\tau} y_t' \rightarrow \Lambda_\tau \quad (37)$$

for arbitrary  $t_0$ , which can be read as a kind of “statistical regularity” of the (future) data. The ergodicity (36) is of course unverifiable in practice as it says something about data which have not been observed yet. Some assumption of this sort about the mechanism generating future data seems however to be necessary to even formulate the iden-

tification problem. We shall call  $\Lambda$  the *true covariance* of the time series  $\{y_t\}$ .

Now, for each  $t \in \mathbb{Z}_+$  define the  $m \times \infty$  matrices

$$\mathbb{y}(t) := [y_t, y_{t+1}, y_{t+2}, \dots] \quad (38)$$

and consider the sequence  $\mathbb{y} := \{\mathbb{y}(t) \mid t \in \mathbb{Z}_+\}$ . We shall make this sequence of semi-infinite matrices into an object isomorphic to the stationary processes  $y$ .

Define the vector space  $\mathbb{Y}$  of scalar semi-infinite real sequences obtained as finite linear combinations of the rows of  $y$ ,

$$\mathbb{Y} := \left\{ \sum a_k' \mathbb{y}(t_k); \quad a_k \in \mathbb{R}^m, \quad t_k \in \mathbb{Z}_+ \right\} \quad (39)$$

This space can be naturally made into an inner product space in the following way.

First, define the bilinear form  $\langle \cdot, \cdot \rangle$  on the generators by letting

$$\begin{aligned} \langle a' \mathbb{y}(k), b' \mathbb{y}(j) \rangle &:= \lim_{N \rightarrow \infty} \frac{1}{N+1} \sum_{t=0}^N a' y_{t+k} y_{t+j}' b \\ &= a' \Lambda_{k-j} b, \end{aligned} \quad (40)$$

for  $a, b \in \mathbb{R}^m$ , and then extend it by linearity to all elements of  $\mathbb{Y}$ .

Let  $\mathbf{a} := \{a_k, k \in \mathbb{Z}_+\}$  be a sequence of vectors  $a_k \in \mathbb{R}^m$ , with compact support in  $\mathbb{Z}_+$ , and let  $\mathbf{a}' := \{a_k'\}$ . A generic element  $\xi$  of the vector space  $\mathbb{Y}$  can be represented as

$$\xi = \sum_k a_k' \mathbb{y}(k) = \mathbf{a}' \mathbb{y}$$

Let us assume that the infinite block-symmetric Toeplitz matrix

$$T = \begin{bmatrix} \Lambda_0 & \Lambda_1 & \dots & \Lambda_k & \dots \\ \Lambda_1' & \Lambda_0 & \Lambda_1 & \dots & \dots \\ \vdots & & \ddots & & \vdots \\ \Lambda_k' & & & \Lambda_0 & \\ \dots & & & & \end{bmatrix} \quad (41)$$

constructed from the “true” covariance sequence  $\{\Lambda_0, \Lambda_1, \dots, \Lambda_k, \dots\}$  of the data, is positive definite. Since the bilinear form (40) on  $\mathbb{Y}$  can be represented by the quadratic form

$$\langle \xi, \eta \rangle = \langle \mathbf{a}' \mathbb{y}, \mathbf{b}' \mathbb{y} \rangle = \sum_{kj} a_k' \Lambda_{k-j} b_j = \mathbf{a}' T \mathbf{b}$$

it can be seen that the bilinear form is nondegenerate (unless  $\Lambda = 0$  identically) and defines a bona-fide inner prod-

uct. As usual one identifies elements whose difference has norm zero (this means  $\langle \xi, \xi \rangle = 0 \Leftrightarrow \xi = 0$ ).

By closing the vector space  $\mathbb{Y}$  with respect to the norm induced by the inner product (40), one obtains a real Hilbert space of semi-infinite sequences which hereafter we shall still denote  $\mathbb{Y}$ . This is the basic data space on which our models will be defined.

The *shift operator*  $U$  operates on the family of semi-infinite matrices (38), by the rule

$$Ua'y(t) = a'y(t+1) \quad t \in \mathbb{Z}, \quad a \in \mathbb{R}^m$$

It is easy to see that  $U$  is a linear map which is isometric with respect to the inner product (40) and can be extended by linearity to all of  $\mathbb{Y}$ . This Hilbert space framework for time series was introduced in [38]. It is shown in this reference that the “stationary Hilbert space”  $(\mathbb{Y}, U)$  is isomorphic to the standard stochastic Hilbert space setup in the  $L^2$  theory of second-order stationary random processes. By virtue of this isomorphism one can formally think of the observed time series (35) as an ergodic sample path of some Gaussian stationary stochastic process  $y$  defined on a true probability space, having covariance matrices equal to the limit  $\Lambda$  of the sum (36) as  $N \rightarrow \infty$ .

Linear functions and operators on the “tail sequences”  $y(t)$  correspond to the same linear functions and operators on the random variables of the process  $y$ . In particular the second-order moments of  $y$  can be computed in terms of the tail sequences  $y$ , by substituting expectations with ergodic limits of the type (40). Since second-order properties are all what matters in our setting, one may even regard the tail sequence  $y$  of (38) as being the *same object* as the underlying stochastic process  $y$ . The usual probabilistic language can be adopted in the present setting provided one identifies real random variables as semi-infinite strings of numbers having the “ergodic property” described at the beginning of this section. The inner product of two semi-infinite strings  $\xi$  and  $\eta$  in  $\mathbb{Y}$  corresponds in particular to the expectation  $\mathbb{E}\{\xi\eta\}$ ; i. e.

$$\langle \xi, \eta \rangle = \mathbb{E}\{\xi\eta\}. \quad (42)$$

This unification of language permits us to carry over in its entirety the geometric theory of stochastic realization derived in the abstract  $L^2$  setting to the time series framework. In particular, the *past* and *future* subspaces of the “processes”  $y$  at time  $t$  are defined as the closure of the linear vector spaces spanned by the relative past or future “random variables”  $y(t)$ , in the metric of the Hilbert space  $\mathbb{Y}$ . The only difference to keep in mind here is

the different interpretation that representation formulas like (8a) have in this context. The equalities involved in the representation

$$\begin{cases} x(t+1) = Ax(t) + Bw(t) \\ y(t) = Cx(t) + Dw(t) \end{cases} \quad (43)$$

are now to be understood in the sense of equalities of elements of  $\mathbb{Y}$ , i. e. as asymptotic equality of sequences in the sense of Cesàro limits. In particular the equality signs in the model (43) do not necessarily imply that the same relations hold for the sample values  $y_t, x_t, w_t$  at a particular instant of time  $t$ . This is in a certain sense similar to the “with probability one” interpretation of the equality sign to be given to the model (43) in case the variables are bona-fide random variables in a probability space.

Consider the orthogonal projection  $\mathbb{E}[\xi | \mathbf{X}]$  of a (row) random variable  $\xi$  onto a subspace  $\mathbf{X}$  of the space  $\mathbb{Y}$ . In the probabilistic  $L^2$  setting this has the well-known interpretation of wide-sense conditional expectation given the random variables in  $\mathbf{X}$  (a true conditional expectation, in the case of Gaussian distributions). In this setting the projection operator has an immediate and useful *statistical* meaning.

Assume for simplicity that  $\mathbf{X}$  is given as the row-space of some matrix of generators  $X$ , then the projection  $\mathbb{E}[\xi | \mathbf{X}]$  has exactly the familiar aspect of the least squares formula expressing the best approximation of the vector  $\xi$  as a linear combination of the rows of  $X$ . For, writing  $\mathbb{E}[\xi | \mathbf{X}]$  to denote the projection expressed (perhaps nonuniquely) in terms of the rows of  $X$ , the classical linear “conditional expectation” formula leads to

$$\mathbb{E}[\xi | \mathbf{X}] = \xi X' [XX']^\# X, \quad (44)$$

which is the universally known “least squares” formula of statistics. The pseudoinverse  $\#$  can be substituted by a true inverse in case the rows of  $X$  are linearly independent.

### The Question of Statistical Efficiency

There are questions about the statistical significance (what are the uncertainty bounds on the parameters and on the estimated transfer functions etc.) which are easily and naturally addressed in the optimization framework but harder to be addressed in the subspace/realization approach to identification. The main difference with the mainstream statistical approach is that the estimation of  $(A, C, \bar{C})$  is not done *directly* by optimizing a likelihood or other distance functions but by just matching second-order moments, i. e. by solving the Eqs. (46).

This way of proceeding can be seen as an instance of *estimation by the method of moments* described in the statistical textbooks e. g. p. 497 in [12], a very old idea used extensively by K. Pearson in the beginning of the last century. The underlying estimation principle is that the parameter estimates should match exactly the sample second-order moments and is close in spirit to the wide-sense setting that we are working in. It does not involve optimality or minimal distance criteria between the “true” and the model distributions. For this reason it is generally claimed in the literature that one should expect better results (in the sense of smaller asymptotic variance of the estimates) by optimization methods. However, a proof that subspace methods can be efficient (under certain conditions which will be too long to report here) has appeared recently [8].

Usable expressions for the asymptotic covariance of the parameter estimates are given in [9]. The arguments in the derivations are, however, far from trivial.

### Subspace Algorithms

Most standard algorithms for subspace identification can be shown to be essentially equivalent to the following three-step procedure (see e. g. [6,38]),

1. The first step is the estimation of a *finite* sequence of sample covariance matrices

$$\{\hat{\Lambda}_0, \hat{\Lambda}_1, \dots, \hat{\Lambda}_\nu\} \quad (45)$$

from the observed data. Since the data are necessarily finite,  $\nu$  must be also finite (and in general small compared to the data length).

2. The second step is identification of a rational model for the covariance sequence. This is a *minimal partial realization* (also called “rational extension”) problem. Given a finite set of experimental covariance data one is asked to find a minimal value of  $n$  and a *minimal triplet* of matrices  $(A, C, \bar{C})$ , of dimensions  $n \times n$ ,  $m \times n$  and  $m \times n$  respectively, such that

$$\hat{\Lambda}_k = CA^{k-1}\bar{C}' \quad k = 1, \dots, \nu. \quad (46)$$

Recall that  $(A, C, \bar{C}')$  is a *minimal triplet*, in the sense of deterministic linear system theory, if the pairs of matrices  $(A, C)$  and  $(A, \bar{C}')$  are an observable and, respectively, a reachable pair, see e. g. [24]. There are well-known algorithms for computing minimal partial realizations (see e. g. [72] for an efficient algorithm) thereby producing “estimates” of the parameters  $(A, C, \bar{C})$  of a minimal realization of a rational spectral density matrix of the process.

3. The third step is to compute, starting from the realization of the rational spectrum estimated in step two, a stationary state-space model. Typically one is interested in the innovation model. This is accomplished by computing the minimal solution  $P_-$  of the Linear Matrix Inequality (16), or, equivalently, the minimal solution of the associated algebraic Riccati equation.

A warning is in order concerning the practical use of these algorithms in that some nontrivial mathematical questions related to positivity of the estimated spectrum are often overlooked in the implementation. This issue is thoroughly discussed in [38] and here we shall just give a short summary.

In determining a minimal triplet  $(A, C, \bar{C})$  interpolating the partial sequence (45) so that  $CA^{k-1}\bar{C}' = \hat{\Lambda}_k$   $k = 1, 2, \dots, \nu$ , we also completely determine the infinite sequence

$$\{\hat{\Lambda}_0, \hat{\Lambda}_1, \hat{\Lambda}_2, \hat{\Lambda}_3, \dots\} \quad (47)$$

by setting  $\hat{\Lambda}_k = CA^{k-1}\bar{C}'$  for  $k = \nu + 1, \nu + 2, \dots$ . This sequence is called a *minimal rational extension* of the finite sequence (45). The attribute “rational” is due to the fact that

$$\begin{aligned} \hat{\Phi}_+(z) &:= \frac{1}{2}\hat{\Lambda}_0 + \hat{\Lambda}_1 z^{-1} + \hat{\Lambda}_2 z^{-2} + \dots \\ &= \frac{1}{2}\hat{\Lambda}_0 + C(zI - A)^{-1}\bar{C}' \end{aligned} \quad (48)$$

is a rational function. In order for (47) to be a bona fide covariance sequence, however, it is necessary that the *infinite* block-Toeplitz matrix obtained by extending the finite matrix

$$T_\nu = \begin{bmatrix} \hat{\Lambda}_0 & \hat{\Lambda}_1 & \hat{\Lambda}_2 & \dots & \hat{\Lambda}_\nu \\ \hat{\Lambda}_1' & \hat{\Lambda}_0 & \hat{\Lambda}_1 & \dots & \hat{\Lambda}_{\nu-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{\Lambda}_\nu' & \hat{\Lambda}_{\nu-1}' & \hat{\Lambda}_{\nu-2}' & \dots & \hat{\Lambda}_0 \end{bmatrix} \quad (49)$$

using the rational extension sequence (47), should be nonnegative definite. Equivalently, the spectral density function corresponding to (47)

$$\hat{\Phi}(z) = \hat{\Lambda}_0 + \sum_{k=1}^{\infty} \hat{\Lambda}_k(z^k + z^{-k}) = \hat{\Phi}_+(z) + \hat{\Phi}_+(z^{-1})' \quad (50)$$

should be nonnegative definite on the unit circle. This is equivalent to the function  $\hat{\Phi}_+(z)$  being *positive real*. Consequently, the partial realization needs to be done subject to the extra constraint of positivity.

The constraint of positivity is a rather tricky one and in some of the methods described in the literature it is disregarded. For this reason some subspace algorithms may fail to provide a positive extension and hence may lead to data  $(A, C, \tilde{C})$  for which there are no solutions of the LMI and hence to totally inconsistent results.

It is important to appreciate the fact that the problem of positivity of the extension has little to do with the noise or sample variability of the covariance data and is present equally well for any *finite* covariance sequence extracted from a true (infinitely long) rational covariance sequence. For there is no guarantee that, even in this idealized situation, the order of a minimal rational extension (47) of the finite covariance subsequence would be sufficiently large to equal the order of the infinite sequence and hence to generate a positive extension. A minimal partial realization may well fail to be positive because its order is too low to guarantee positivity. It is shown in [38] that there are relatively “large” sets of data (45) for which this happens and the matrix  $A$  may even fail to be stable [7].

### Future Directions

Because of the pervasive presence of fast digital computers in modern society, we are witnessing a sharp change of paradigms in the analysis, design, prediction and control of both man made and natural (say biological, economical etc) systems. Because of the tremendous computing capabilities at our disposal nowadays, stochastic modeling, realization and identification have become an essential ingredient of modern applied sciences. In particular modeling and simulation of complex technological, biological, economic, and environmental systems is becoming more and more essential for understanding the behavior and for prediction and control of these systems. New problems continually arise and much work remains to be done.

We just mention here problems where there is a need of modeling the effect of *input* or *decision variables*, possibly in the presence of feedback. Such *Input-Output* models have important applications in many applied areas such as control of industrial processes, econometrics, etc. There is a large body of results concerning stochastic realization and subspace identification of systems with inputs which we could not discuss in this article, see [10,11,26,27,53,54,55,64,65]. The area is important for applications and still needs research. Modeling and realization of finite state processes and of certain classes of highly non-Gaussian signals is needed in diverse applications such as source coding, image processing, and computer vision.

## Bibliography

### Primary Literature

1. Akaike H (1975) Markovian representation of stochastic processes by canonical variables. *SIAM J Control* 13:162–173
2. Akaike H (1974) Stochastic Theory of Minimal Realization. *IEEE Trans Autom Control* AC-19(6):667–674
3. Anderson BDO (1969) The inverse problem of Stationary Covariance Generation. *J Stat Phys* 1(1):133–147
4. Anderson BDO (1967) An algebraic solution to the spectral factorization problem. *IEEE Trans Autom Control* AC-12:410–414
5. Anderson BDO (1999) The realization problem for Hidden Markov Models. *Math Control Signals Syst* 12:80–120
6. Aoki M (1990) *State Space Modeling of Time Series*, 2nd edn. Springer, New York
7. Byrnes CI, Lindquist A (1982) The stability and instability of partial realizations. *Syst Control Lett* 2:2301–2312
8. Bauer D (2005) Comparing the CCA Subspace Method to Pseudo Maximum Likelihood Methods in the case of No Exogenous Inputs. *J Time Ser Anal* 26:631–668
9. Chiuso A, Picci G (2004) The Asymptotic Variance of Subspace Estimates. *J Econom* 118(1–2):257–291
10. Chiuso A, Picci G (2004) On the Ill-conditioning of subspace identification with inputs. *Automatica* 40(4):575–589
11. Chiuso A, Picci G (2005) Consistency Analysis of some Closed-loop Subspace Identification Methods. *Automatica* 41:377–391, special issue on System Identification
12. Cramér H (1949) *Mathematical Methods of Statistics*. Princeton University Press, Princeton
13. Desai UB, Pal D, Kirkpatrick RD (1985) A Realization Approach to Stochastic Model Reduction. *Int J Control* 42:821–838
14. Doob JL (1953) *Stochastic Processes*. Wiley, New York
15. Faurre P (1969) Identification par minimisation d’une représentation Markovienne de processus aleatoires. Symposium on Optimization, Nice 1969. *Lecture Notes in Mathematics*, vol 132. Springer, New York
16. Faurre P, Chataigner P (1971) Identification en temps reel et en temps differeé par factorisation de matrices de Hankel. French-Swedish colloquium on process control, IRIA Roquencourt
17. Faurre P (1973) Realisations Markovienne de processus stationnaires. Report de Recherche no 13, INRIA, Roquencourt
18. Ferrante A, Picci G, Pinzoni S (2002) Silverman algorithm and the structure of discrete-time stochastic systems. *Linear Algebra Appl* (special issue on systems and control) 351–352:219–242
19. Finesso L, Spreij PJC (2002) Approximate realization of finite Hidden Markov Chains. *Proceedings of the 2002 IEEE Information Theory Workshop, Bangalore*, pp 90–93
20. Ford GW, Kac M, Mazur P (1965) Statistical mechanics of assemblies of coupled oscillators. *J Math Phys* 6:504–515
21. Kalman RE (1963) On a new Criterion of Linear Passive Systems. In: *Proc. of the First Allerton Conference*. University of Illinois, Urbana Ill, Nov 1963, pp 456–470
22. Kalman RE (1963) Lyapunov Functions for the problem of Lur’e in Automatic Control. *Proc Natl Acad Sci* 49:201–205
23. Kalman RE (1965) Linear stochastic filtering: reappraisal and outlook. In: *Proc. Symposium on System Theory*, Politechnic Institute of Brooklin, pp 197–205
24. Kalman RE, Falb PL, Arbib MA (1969) *Topics in Mathematical Systems Theory*. McGraw-Hill, New York

25. Karatzas I, Shreve S (1987) Brownian motion and stochastic calculus. Springer, New York
26. Katayama T, Picci G (1999) Realization of Stochastic Systems with Exogenous Inputs and Subspace Identification Methods. *Automatica* 35(10):1635–1652
27. Katayama T (2005) Subspace methods for System Identification. Springer, New York
28. Kintchine A (1934) Korrelationstheorie Stationäre Stochastischen Prozessen. *Math Annalen* 109:604–615
29. Lax PD, Phillips RS (1967) Scattering Theory. Academic Press, New York
30. Levy P (1948) Processus stochastiques et Mouvement Brownien. Gauthier-Villars, Paris
31. Lewis JT, Thomas LC (1975) How to make a heat bath. In: Arthurs AM (ed) *Functional Integration and its Applications*. Clarendon, Oxford
32. Lewis JT, Maassen H (1984) Hamiltonian models of classical and quantum stochastic processes. In: Accardi L, Frigerio A, Gorini V (eds) *Quantum Probability and Applications to the Quantum Theory of Irreversible processes*. Springer Lecture Notes in Mathematics, vol 1055. Springer, New York, pp 245–276
33. Lindquist A, Picci G, Ruckebusch G (1979) On minimal splitting subspaces and Markovian representation. *Math Syst Theory* 12:271–279
34. Lindquist A, Picci G (1979) On the stochastic realization problem. *SIAM J Control Optim* 17:365–389
35. Lindquist A, Picci G (1985) Realization theory for multivariate stationary Gaussian processes. *SIAM J Control Optim* 23:809–857
36. Lindquist A, Pavon M (1984) On the structure of state space models of discrete-time vector processes. *IEEE Trans Autom Control* AC-29:418–432
37. Lindquist A, Picci G (1991) A geometric approach to modelling and estimation of linear stochastic systems. *J Math Syst Estim Control* 1:241–333
38. Lindquist A, Picci G (1996) Canonical correlation analysis approximate covariance extension and identification of stationary time series. *Automatica* 32:709–733
39. Lindquist A, Michaletzky G, Picci G (1995) Zeros of Spectral Factors, the Geometry of Splitting Subspaces, and the Algebraic Riccati Inequality. *SIAM J Control Optim* 33:365–401
40. Lindquist A, Picci G (1996) Geometric Methods for State Space Identification. In: *Identification, Adaptation, Learning*. (Lectures given at the NATO-ASI School, From Identification to Learning, Como, Italy, Aug 1994. Springer
41. Lindquist A, Picci G, Linear Stochastic Systems: A Geometric Approach. (in preparation)
42. Ljung L (1987) *System Identification – Theory for the User*. Prentice-Hall, New York
43. Martin-Löf A (1979) *Statistical Mechanics and the foundations of Thermodynamics*. Lecture Notes in Physics, vol 101. Springer, New York
44. Moonen M, De Moor B, Vanderberghe L, Vandewalle J (1989) On- and Off-Line Identification of Linear State-Space Models. *Int J Control* 49:219–232
45. Taylor TJ, Pavon M (1988) A solution of the nonlinear stochastic realization problem. *Syst Control Lett* 11:117–121
46. Picci G (1978) On the internal structure of finite-state stochastic processes. In: Mohler R, Ruberti A (eds) *Recent developments in Variable Structure Systems*. Springer Lecture Notes in Economics and Mathematical Systems. vol 162. Springer, New York, pp 288–304
47. Picci G, Pinzoni S (1994) Acausal Models and Balanced realizations of stationary processes. *Linear Algebra Appl* (special issue on Systems Theory) 205–206:957–1003
48. Picci G (1986) Application of stochastic realization theory to a fundamental problem of statistical physics. In: Byrnes CI, Lindquist A (eds) *Modeling, Identification and Robust Control*. North Holland, Amsterdam
49. Picci G, Taylor TJ (1990) Stochastic aggregation of linear Hamiltonian systems with microcanonical distribution. In: Kaashoek MA, van Schuppen JH, Ran ACM (eds) *Realization and modeling in System Theory*. Birkhäuser, Boston, pp 513–520
50. Picci G (1992) Markovian representation of linear Hamiltonian systems. In: Guerra F, Loffredo MI, Marchioro C (eds) *Probabilistic Methods in Mathematical Physics*. World Sciedntific, Singapore, pp 358–373
51. Picci G (1991) Stochastic realization theory. In: Antoulas A (ed) *Mathematical System Theory*. Springer, New York
52. Picci G, Taylor TSJ (1992) Generation of Gaussian Processes and Linear Chaos. In: *Proc 31st IEEE Conf on Decision and Control*. Tucson, pp 2125–2131
53. Picci G, Katayama T (1996) Stochastic realization with exogenous inputs and “Subspace Methods” Identification. *Signal Process* 52(2):145–160
54. Picci G (1996) Geometric methods in Stochastic Realization and System Identification. *CWI Q* (invited paper) 9:205–240
55. Picci G (1997) Oblique Splitting Subspaces and Stochastic Realization with Inputs. In: Helmke U, Prätzel-Wolters D, Zerz E (eds) *Operators, Systems and Linear Algebra*. Teubner, Stuttgart, pp 157–174
56. Popov VM (1964) Hyperstability and Optimality of Automatic Systems with several Control functions. *Rev Romaine Sci Tech Ser Electrotech Energ* 9:629–690
57. Rozanov NI (1963) *Stationary Random Processes*. Holden-Day, San Francisco
58. Ruckebusch G (1976) Représentations markoviennes de processus gaussiens stationnaires. *Comptes Rendues Acad Sci Paris Series A* 282:649–651
59. Ruckebusch G (1978) A state space approach to the stochastic realization problem. *Proc 1978 IEEE Intern Symp Circuits and Systems*, pp 972–977
60. Ruckebusch G (1978) Factorisations minimales de densités spectrales et représentations markoviennes. In: *Proc 1re Colloque AFCET-SMF, Palaiseau*
61. Söderström T, Stoica P (1989) *System Identification*. Prentice-Hall, New York
62. van Schuppen JH (1989) Stochastic realization problems. In: Nijmeijer H, Schumacher JM (eds) *Three decades of mathematical system theory. Lecture Notes in Control and Information Sciences*, vol 135. Springer, New York, pp 480–532
63. Van Overschee P, De Moor B (1993) Subspace algorithms for the stochastic identification problem. *Automatica* 29:649–660
64. Van Overschee P, De Moor B (1994) N4SID: Subspace algorithms for the identification of combined deterministic-stochastic systems. *Automatica* 30:75–93
65. Verhaegen M (1994) Identification of the deterministic part of MIMO State Space Models given in Innovations form from Input-Output data. *Automatica* 30:61–74
66. Vidyasagar M, (Hidden) Markov Models: Theory and Applications to Biology. Princeton University Press, Princeton



67. Wiener N (1930) Generalized Harmonic Analysis. *Acta Math* 55:117–258
68. Wiener N (1933) Generalized Harmonic Analysis. In: *The Fourier Integral and Certain of its Applications*. Cambridge University Press, Cambridge
69. Wiener N (1949) *Extrapolation, Interpolation and Smoothing of Stationary Time Series*. The M.I.T. Press, Cambridge
70. Willems JC (1971) Least Squares Stationary Optimal Control and the Algebraic Riccati Equation. *IEEE Trans Autom Control* AC-16:621–634
71. Yakubovich VA (1963) Absolute stability of Nonlinear Control Systems in critical case: part I and II. *Avtom Telemekh* 24:293–303, 717–731
72. Zeiger HP, McEwen AJ (1974) Approximate linear realization of given dimension via Ho's algorithm. *IEEE Trans Autom Control* AC-19:153

### Books and Reviews

- Doob JL (1942) The Brownian movement and stochastic equations. *Ann Math* 43:351–369
- Faure P, Clerget M, Germain F (1979) *Opérateurs Rationnels Positifs*. Dunod, Paris
- Kalman RE (1981) Realization of covariance sequences. *Proc Toeplitz Memorial Conference*, Tel Aviv, Birkhäuser
- Nelson E (1967) *Dynamical theories of Brownian motion*. Princeton University press, Princeton
- Picci G (1976) Stochastic realization of Gaussian processes. *Proc IEEE* 64:112–122
- Picci G (1992) Stochastic model reduction by aggregation. In: Isidori A, Tarn TJ (eds) *Systems Models and Feedback: Theory and Applications*. Birkhäuser Verlag, Basel, pp 169–177

## Stochastic Processes

ALAN J. MCKANE

Theory Group, School of Physics and Astronomy,  
University of Manchester, Manchester, UK

### Article Outline

Glossary  
[Definition of the Subject](#)  
[Introduction](#)  
[Markov Chains](#)  
[The Master Equation](#)  
[The Fokker–Planck Equation](#)  
[Stochastic Differential Equations](#)  
[Path Integrals](#)  
[System Size Expansion](#)  
[Future Directions](#)  
[Bibliography](#)

### Glossary

**Fokker–Planck equation** A partial differential equation of the second order for the time evolution of the proba-

bility density function of a stochastic process. It resembles a diffusion equation, but has an extra term which represents the deterministic aspects of the process.

**Langevin equation** A stochastic differential equation of the simplest kind: linear and with an additive Gaussian white noise. Introduced by Langevin [30] in 1908 to describe Brownian motion; many stochastic differential equations in physics go by this name.

**Markov process** A stochastic process in which the current state of the system is only determined from its state in the immediate past, and not by its entire history.

**Markov chain** A Markov process where both the states and the time are discrete and where the process is stationary.

**Master equation** The equation describing a continuous-time Markov chain.

**Stochastic process** A sequence of stochastic variables. This sequence is usually a time-sequence, but could also be spatial.

**Stochastic variable** A random variable. This is a function which maps outcomes to numbers (real or integer).

### Definition of the Subject

The most common type of stochastic process comprises of a set of random variables  $\{x(t)\}$ , where  $t$  represents the time which may be real or integer valued. Other types of stochastic process are possible, for instance when the stochastic variable depends on the spatial position  $\mathbf{r}$ , as well as, or instead of,  $t$ . Since in the study of complex systems we will predominantly be interested in applications relating to stochastic dynamics, we will suppose that it depends only on  $t$ . One of the earliest investigations of a stochastic process was carried out by Bachelier [2], who used the idea of a random walk to analyze stock market fluctuations. The problem of a random walk was more generally discussed by Pearson [42], and applied to the investigation of Brownian motion by Einstein [8,9], Smoluchowski [54] and Langevin [30]. The example of a random walk illustrates the fact that in addition to time being discrete or continuous, the stochastic variable itself can be discrete (for instance, the walker moves with fixed step size in one dimension) or continuous (for instance, the velocity of a Brownian particle). The modeling of the process may lead to an equation for the stochastic variable, such as a stochastic differential equation, or for an equation which predicts how the probability density function (pdf) for the stochastic variable changes in time. Stochastic processes are ubiquitous in the physical, biological and social sciences; they may come about through the perception of

very complicated processes being essentially random (the toss of a coin, roll of a die, birth or death of individuals in populations), the inclusion of diverse and poorly characterized effects external to the system under consideration, or thermal fluctuations, among others.

## Introduction

Deterministic dynamical processes are typically formulated as a set of rules which allow for the state of the system at time  $t + 1$  (or  $t + \delta t$ ) to be found from the state of the system at time  $t$ . By contrast, for stochastic systems, we can only specify the probability of finding the system in a given state. If this only depends on the state of the system at the previous time step, but not those before this, the stochastic process is said to be Markov. Fortunately many stochastic processes are Markovian to a very good approximation, since the theory of non-Markov processes is considerably more complicated than Markov processes and much less well developed. In this article we will deal almost exclusively with Markov processes.

The mathematical definition of a Markov process follows from the definition of the hierarchy of pdfs for a given process. This involves the joint pdfs  $P(x_1, t_1; x_2, t_2; \dots; x_n, t_n)$ , which are the probability that the system is in state  $x_1$  at time  $t_1$ , state  $x_2$  at time  $t_2, \dots$ , and state  $x_n$  at time  $t_n$ , and also the conditional pdfs  $P(x_1, t_1; \dots; x_m, t_m | x_{m+1}, t_{m+1}; \dots; x_n, t_n)$ , which are the probability that the system is in state  $x_1$  at time  $t_1, \dots, x_m$  at time  $t_m$ , given that it was in state  $x_{m+1}$  at time  $t_{m+1}, \dots, x_n$  at time  $t_n$ . These pdfs are all non-negative and normalizable, and relations exist between them due to symmetry and reduction (integration over some of the state variables). Nevertheless, for a general non-Markov process, a whole family of these pdfs will be required to specify the process. On the other hand, for a Markov process the history of the system, apart from the immediate past, is forgotten, and so  $P(x_1, t_1; \dots; x_m, t_m | x_{m+1}, t_{m+1}; \dots; x_n, t_n) = P(x_1, t_1; \dots; x_m, t_m | x_{m+1}, t_{m+1})$ . A direct consequence of this is that the whole hierarchy of pdfs can be determined from only two of them:  $P(x, t)$  and  $P(x, t | x', t')$ . The hierarchy of defining equations then collapses to only two:

$$P(x_2, t_2) = \int dx_1 P(x_2, t_2 | x_1, t_1) P(x_1, t_1) \quad (1)$$

and

$$P(x_3, t_3 | x_1, t_1) = \int dx_2 P(x_3, t_3 | x_2, t_2) P(x_2, t_2 | x_1, t_1), \quad t_1 < t_2 < t_3. \quad (2)$$

The pdf  $P(x, t | x', t')$  is referred to as the transition probability and Eq. (2) as the Chapman–Kolmogorov equation. While the pdfs for a Markov process must obey Eqs. (1) and (2), the converse also holds: any two non-negative functions  $P(x, t)$  and  $P(x, t | x', t')$  which satisfy Eqs. (1) and (2), uniquely define a Markov process.

We will begin our discussion in Sect. “[Markov Chains](#)” with what is probably the simplest class of Markov processes: the case when both the state space and time are discrete. These are called Markov chains and were first investigated, for a finite number of states, by Markov [32] in 1906. The extension to an infinite number of states was carried out by Kolmogorov [27] in 1936. If time is continuous, an analogous formalism may be developed, which will be discussed in Sect. “[The Master Equation](#)”. In physics the equation describing the time evolution of the pdf in this case is called the master equation and was introduced by Pauli [41] in 1928, in connection with the approach to equilibrium for quantum systems, and also by Nord-sieck, Lamb and Uhlenbeck [39] in 1940, in connection with fluctuations in cosmic ray physics. The term “master equation” refers to that fact that many of the quantities of interest can be derived from this equation. The connection with previous work on Markov processes was clarified by Siegert [49] in 1949.

In many instances when the master equation cannot be solved exactly, it is useful to approximate it by a rather coarser description of the system, known as the Fokker–Planck equation. This approach will be discussed in Sect. “[The Fokker–Planck Equation](#)”. This equation was used in its linear form by Rayleigh [44], Einstein [8,9], Smoluchowski [54,55], and Fokker [14], but it was Planck [43] who derived the general nonlinear form from a master equation in 1917, and Kolmogorov who made the procedure rigorous in 1931. All the descriptions which we have mentioned so far have been based on the time evolution of the pdfs. An alternative specification is to give the time evolution of the stochastic variables themselves. This will necessarily involve random variables appearing in the equations describing this evolution, and they will therefore be *stochastic differential equations*. The classic example is the Langevin equation [30] used to describe Brownian motion. This equation is linear and can therefore be solved exactly. The Langevin approach, and its relation to the Fokker–Planck equation is described in Sect. “[Stochastic Differential Equations](#)”.

A good summary of the understanding of stochastic processes that had been gained by the mid-1950s is given in the book edited by Wax. This covers the basics of the subject, and what is discussed in the first six sections of this article. The article by Chandrasekhar [4], first

published in 1943, and reprinted in Wax, gives an extensive bibliography of stochastic problems in physics before 1943. Since then the applications of the subject have grown enormously, and the equations modeling these systems have correspondingly become more complex. We illustrate some of the procedures which have been developed to deal with these equations in the next two sections. In Sect. “[Path Integrals](#)” we discuss how the path-integral formalism may be applied to stochastic processes and in Sect. “[System Size Expansion](#)” we describe how master equations can be analyzed when the size of the system is large. We end with a look forward to the future in Sect. “[Future Directions](#)”.

## Markov Chains

The simplest version of the Markov process is when both the states and the time are discrete, and when the stochastic process is *stationary*. When the states are discrete we will denote them by  $n$  or  $m$ , rather than  $x$ , which we reserve for continuous state variables. In this notation the two Eqs. (1) and (2) governing Markov processes read

$$P(n_2, t_2) = \sum_{n_1} P(n_2, t_2 | n_1, t_1) P(n_1, t_1) \quad (3)$$

$$P(n_3, t_3 | n_1, t_1) = \sum_{n_2} P(n_3, t_3 | n_2, t_2) P(n_2, t_2 | n_1, t_1), \quad t_1 < t_2 < t_3. \quad (4)$$

A stationary process is one in which the conditional pdf  $P(n, t | n', t')$  only depends on the time difference  $(t - t')$ . For such processes, when time is discrete so that  $t = t' + 1, t' + 2, \dots$ , we may write  $P(n, t' + k | n', t')$  as  $p_{n n'}^{(k)}$ . The most elementary form of the Chapman–Kolmogorov equation (4) may then be expressed as

$$p_{n m}^{(2)} = \sum_{n'} p_{n n'}^{(1)} p_{n' m}^{(1)}. \quad (5)$$

This corresponds to the matrix multiplication of  $p^{(1)}$  with itself, and therefore  $p^{(2)}$  is simply  $(p^{(1)})^2$ . In the same way  $p^{(k)} = (p^{(1)})^k$ , and from now on we drop the superscript on  $p^{(1)}$  and denote the matrix by  $\mathcal{P}$ . The entries of  $\mathcal{P}$  are non-negative, with the sum of entries in each column being equal to unity, since

$$\sum_n p_{n n'} = \sum_n P(n, t + 1 | n', t) = 1. \quad (6)$$

Such matrices are called *stochastic matrices*. From Eq. (5) it is clear that  $\mathcal{P}^2$  is also a stochastic matrix, and by induction it follows that  $\mathcal{P}^k$  is a stochastic matrix if  $\mathcal{P}$  is.

The other defining relation for a Markov process, Eq. (3), now becomes

$$P(n, t + 1) = \sum_{n'} p_{n n'} P(n', t). \quad (7)$$

This relation defines a *Markov chain*. It has two ingredients: the probability that the system is in state  $n$  at time  $t$ ,  $P(n, t)$  – which is usually what we are trying to determine, and the stochastic matrix with entries  $p_{n n'}$  which gives the probabilities of transitions from the state  $n'$  to the state  $n$ . The transition probabilities are typically given; they define the model. Note that in many texts the probability of making a transition from  $n'$  to  $n$  is written as  $p_{n' n}$ , not  $p_{n n'}$ . If we write  $P(n, t)$  as a vector  $\mathbf{P}(t)$ , then we may write Eq. (7) as  $\mathbf{P}(t + 1) = \mathcal{P}\mathbf{P}(t)$ . Therefore,

$$\mathbf{P}(t) = \mathcal{P}\mathbf{P}(t - 1) = \mathcal{P}^2\mathbf{P}(t - 2) = \dots = \mathcal{P}^t\mathbf{P}(0), \quad (8)$$

and so if the initial state of the system  $\mathbf{P}(0)$  is given, then we can find the state of the system at time  $t$  ( $\mathbf{P}(t)$ ) by matrix multiplication by the  $t$ th power of the transition matrix.

## Examples of Markov Chains

1. A *one-dimensional random walk*. The most widely known example of a Markov chain is a random walk on the real axis, where the walker takes single steps between integers on the line. The simplest version is where the walker has to move during every time interval:

$$p_{n n'} = \begin{cases} p, & \text{if } n = n' + 1 \\ q, & \text{if } n = n' - 1 \\ 0, & \text{otherwise,} \end{cases} \quad (9)$$

where  $p + q = 1$ . There are many variants. For instance, the walker could have a non-zero probability of staying put, in which case  $p_{n n} = r$ , with  $p + q + r = 1$ . The walk could be heterogeneous, in which case  $p$  and  $q$  (and  $r$ ), could depend on  $n'$ . If there are boundaries, the boundary conditions have to be given. The most common two are *absorbing boundaries* defined by

$$\begin{aligned} p_{n+1 n} &= p, & p_{n-1 n} &= q, & (n = 2, \dots, N - 1) \\ p_{11} &= 1, & p_{NN} &= 1, \\ p_{n n'} &= 0, & \text{otherwise,} \end{aligned} \quad (10)$$

and reflecting boundaries defined by

$$\begin{aligned} p_{n+1n} &= p, \quad p_{n-1n} = q, \quad (n = 2, \dots, N-1) \\ p_{21} &= p, \quad p_{N-1N} = q, \\ p_{11} &= q, \quad p_{NN} = p, \\ p_{nn'} &= 0, \quad \text{otherwise.} \end{aligned} \quad (11)$$

With absorbing boundaries (10), if we reach the state 1 or  $N$ , we can never leave it, since we stay there with probability 1. When the boundary is reflecting, we can never move beyond it; the only options are to move back towards the other boundary, or stay put. Well known examples of absorbing boundaries include the gambler's ruin problem, where a gambler bets a given amount against the house at each time step and can win with a probability  $p$ . Here the absorbing boundary is situated at  $n = 0$ . Eventually he will arrive at the state  $n = 0$ , where he has no money left, and so cannot re-enter the game. Birth/death processes will also have absorbing states at  $n = 0$ : if  $n$  is the number of individuals at a given time, and  $p$  is the probability of a birth and  $q$  of a death, then if there are no individuals left ( $n = 0$ ), none can be born. This condition will be automatically applied if the transition probabilities are proportional to the number of individuals in the population.

2. *The Ehrenfest urn* This Markov chain was introduced by the Ehrenfests [7] in 1907, to illustrate the approach to equilibrium in a gas. Two containers,  $A$  and  $B$ , contain molecules of the same gas, the sum of the number of molecules in  $A$  and  $B$  being fixed to be  $N$ . At time  $t$  a molecule is removed at random from the containers and put into the other. If  $n'$  is the number of molecules in container  $A$  at a certain time, then at the next time step the transition probabilities will be:

$$p_{nn'} = \begin{cases} \frac{n'}{N}, & \text{if } n' = n + 1 \\ (1 - \frac{n'}{N}), & \text{if } n' = n - 1 \\ 0, & \text{otherwise.} \end{cases} \quad (12)$$

This is clearly a heterogeneous random walk of the type (9), and another interpretation of this model is as a random walk, but with a central force.

The most frequently asked question concerning Markov chains is: what is their eventual fate; how does the system behave at large time? Clearly if it tends towards a non-trivial stationary state,  $P_{\text{st}}(n)$ , then from Eq. (7):

$$P_{\text{st}}(n) = \sum_{n'} p_{nn'} P_{\text{st}}(n'), \quad (13)$$

and so  $P_{\text{st}}(n)$  is a right eigenvector of  $\mathcal{P}$  with unit eigenvalue. It follows from the properties of a general stochastic

matrix that the eigenvalues of a stochastic matrix are such that  $|\lambda| \leq 1$  [15]. Furthermore every stochastic matrix has an eigenvalue equal to 1, however it may not be simple – there may be a multiplicity of unit eigenvalues. The classification of Markov chains can be used to decide which of these possibilities is the case. For example, Markov chains may be reducible or irreducible, and states recurrent or transient. We shall not discuss this in detail; Feller [10] gives a clear account of this classification and Cox and Miller [5] explore the consequences for the nature of the eigenvalues. Instead we will examine a specific example, that of the Ehrenfest urn introduced above, and focus on the explicit calculation of the eigenvalues and eigenvectors in that case.

Suppose that  $\Psi^{(k)}$  and  $\Theta^{(k)}$  are the right- and left-eigenvectors of  $\mathcal{P}$ , respectively, corresponding to the eigenvalue  $\lambda^{(k)}$ , so that  $\Theta^{(k)} \cdot \Psi^{(\ell)} = \delta_{k\ell}$ . Then, in general, and for the Ehrenfest urn in particular,

$$\mathcal{P}_{nn'}^t = \sum_{k=0}^N \psi_n^{(k)} \left( \lambda^{(k)} \right)^t \theta_{n'}^{(k)}, \quad (14)$$

where  $\psi_n^{(k)}$  is the  $n$ th component of the vector  $\Psi^{(k)}$  and similarly for the left-eigenvector. The eigenvalues and eigenvectors for the Ehrenfest urn can be found exactly (Kac [23]; see also Krafft and Schaefer [28]). The eigenvalues are  $\lambda^{(k)} = 1 - (2k/N)$ ,  $k = 0, \dots, N$ . Thus in this case there is a single eigenvalue  $\lambda = 1$ . The corresponding right-eigenvector, which gives the stationary state, is a binomial distribution:

$$P_{\text{st}}(n) = \psi_n^{(0)} = \frac{N!}{n!(N-n)!} \frac{1}{2^N}. \quad (15)$$

The left-eigenvector is  $\theta_n^{(0)} = 1$  for all  $n$ . The other eigenvectors have the form  $\theta_n^{(k)} = a_{kn}$  and  $\psi_n^{(k)} = a_{kn} P_{\text{st}}(n)$ , where the  $a_{kn}$  are the Krawtchouk polynomials [1]. Clearly,  $a_{0n} = 1$ , and the first non-trivial polynomial is  $a_{1n} = \sqrt{N}[1 - (2n/N)]$ . Therefore, for a suitable choice of initial conditions and using Eq. (8), the large  $t$  behavior of the Ehrenfest urn can be found from

$$\mathcal{P}_{nn'}^t \approx P_{\text{st}}(n) \left\{ 1 + c_{nn'} \left( 1 - \frac{2}{N} \right)^t \right\}, \quad (16)$$

where  $c_{nn'} = N[1 - (2n/N)][1 - (2n'/N)]$ .

### The Master Equation

The master equation is a Markov chain in the limit where time is continuous. To derive it we will assume that the states are discrete (the derivation is essentially identical

if they are continuous) and write down the Chapman–Kolmogorov equation (4) in the form:

$$P(n, t + \Delta t | n_0, t_0) = \sum_{n'} P(n, t + \Delta t | n', t) P(n', t | n_0, t_0). \quad (17)$$

We consider only stationary processes, so that we may take  $t_0 = 0$  without loss of generality and  $P(n, t + \Delta t | n', t)$  is independent of  $t$ . We now assume that

$$P(n, t + \Delta t | n', t) = \begin{cases} 1 - \kappa(n)\Delta t + o(\Delta t), & \text{if } n = n' \\ T(n|n')\Delta t + o(\Delta t), & \text{if } n \neq n', \end{cases} \quad (18)$$

where  $o(\Delta t)$  means that  $o(\Delta t)/\Delta t$  tends to zero as  $\Delta t \rightarrow 0$ . This is reasonable: after a very short time the transition probability to stay put is unity minus a term of order  $\Delta t$  and the transition probabilities to move to any other state is of order  $\Delta t$ , but this is still an additional assumption on the process. The quantity  $T(n|n')$  is the transition rate, and is only defined for  $n \neq n'$ . Since  $\sum_n P(n, t + \Delta t | n', t) = 1$  for all  $n'$ , we have that

$$\kappa(n') = \sum_{n \neq n'} T(n|n'). \quad (19)$$

Substituting Eq. (18) into Eq. (17), and making use of Eq. (19) we find that

$$\begin{aligned} & \frac{P(n, t + \Delta t | n_0, 0) - P(n, t | n_0, 0)}{\Delta t} \\ &= \sum_{n' \neq n} [T(n|n')P(n', t | n_0, 0)] \\ & \quad - P(n, t | n_0, 0) \sum_{n' \neq n} [T(n'|n)] + \frac{o(\Delta t)}{\Delta t}. \end{aligned} \quad (20)$$

Taking the limit  $\Delta t \rightarrow 0$  gives the master equation for how the probability of finding the system in state  $n$  at time  $t$  changes with time:

$$\frac{dP(n, t)}{dt} = \sum_{n' \neq n} T(n|n') P(n', t) - \sum_{n' \neq n} T(n'|n) P(n, t). \quad (21)$$

We have dropped the initial conditions, assuming that they are understood. It should be noticed that an analogous analysis starting from Eq. (3), rather than Eq. (4), may be carried out, leading to identical equations for

$P(n, t | n_0, 0)$  and  $P(n, t)$ . If the state space is continuous the master equation reads

$$\frac{\partial P(x, t)}{\partial t} = \int dx' [T(x|x') P(x', t) - T(x'|x) P(x, t)]. \quad (22)$$

In most applications transitions only take place between states whose label differs by one. That is,  $T(n|n')$  and  $T(n'|n)$  are zero unless  $n' = n + 1$  and  $n' = n - 1$ . These are called *one-step processes*. For such processes the master equation takes the simpler form

$$\begin{aligned} \frac{dP(n, t)}{dt} &= T(n|n+1)P(n+1, t) \\ & \quad + T(n|n-1)P(n-1, t) \\ & \quad - [T(n-1|n) + T(n+1|n)] P(n, t). \end{aligned} \quad (23)$$

For simplicity let us write

$$g_n = T(n+1|n) \quad \text{and} \quad r_n = T(n-1|n), \quad (24)$$

then the master equation may be written as

$$\begin{aligned} \frac{dP(n, t)}{dt} &= r_{n+1}P(n+1, t) + g_{n-1}P(n-1, t) \\ & \quad - [r_n + g_n] P(n, t). \end{aligned} \quad (25)$$

## Examples of Master Equations

1. *The simple birth–death process.* For a population of simple organisms, for example a colony of bacteria, it might be reasonable to assume that the rate of birth of new bacteria is proportional to the number present at that time, and similarly for the rate of death. This is clearly a Markov process with  $g_n = bn$  and  $r_n = dn$ , where  $b$  and  $d$  are rate constants. A variant is to include “immigrants” coming into the population from the outside at a constant rate  $c$ , so that  $g_n = bn + c$ .

In this example  $g_n$  and  $r_n$  are linear in  $n$ . Such *linear one-step processes* can be solved by the introduction of the generating function  $F(z, t) = \sum_n P(n, t)z^n$ . This converts the master equation (which is a differential-difference equation) into a partial differential equation for  $F(z, t)$  which can be solved if the process is linear. The simplest case of a pure death process ( $b = c = 0$  in the above) can illustrate the general procedure. By rescaling the time ( $t = \tau/d$ ), we may write the master equation in the very simple form

$$\frac{dP(n, \tau)}{d\tau} = (n+1)P(n+1, \tau) - nP(n, \tau).$$



Multiplying this equation by  $z^n$  and summing over all  $n \geq 0$  gives

$$\begin{aligned} & \frac{\partial}{\partial \tau} \left\{ \sum_{n=0}^{\infty} z^n P(n, \tau) \right\} \\ &= \sum_{n=0}^{\infty} (n+1) z^n P(n+1, \tau) - \sum_{n=1}^{\infty} n z^n P(n, \tau) \\ &= \sum_{m=1}^{\infty} m z^{m-1} P(m, \tau) - z \sum_{n=1}^{\infty} n z^{n-1} P(n, \tau), \end{aligned}$$

that is,

$$\frac{\partial F}{\partial \tau} = (1-z) \frac{\partial F}{\partial z}.$$

A change of variable to  $\xi = (1-z)e^{-\tau}$  and  $\eta = \tau$  shows  $F$  to be a function of  $\xi$  only:  $F(z, \tau) = \phi([1-z]e^{-\tau})$ . The function  $\phi$  may be determined from the initial condition. For instance, if  $P(n, 0) = \delta_{nN}$ , then  $F(z, 0) = z^N$  and so  $\phi(\xi) = (1-\xi)^N$ . This gives the solution for  $F$  to be

$$F(z, \tau) = [(1 - e^{-\tau}) + ze^{-\tau}]^N.$$

In this case  $F$  can easily be expanded as a power series in  $z$ , and the  $P(n, \tau)$  read off, but even if this is difficult, the moments of the distribution can be readily found by differentiation with respect to  $z$  and then setting  $z = 1$ . It should now be clear why  $F$  is called a generating function. In the general case the partial differential equation for  $F$  may be solved by standard methods [50]. The solution for a birth–death process without immigration is given in the book by Reichl [45]. The solution with immigration was first given by Kendall [25], who also introduced the technique of the generating function as a method of solution of the master equation.

2. *The Moran model of genetic drift.* Stochastic processes occur extensively in population genetics. The simplest, and most widely known, is a model of genetic drift introduced by Fisher [13] and Wright [58], in which a population of individuals in generation  $t$  mate randomly to produce the new generation  $t+1$ . We assume, for simplicity, that each individual has only one gene of a particular type, and that this may exist in one of two forms (alleles) denoted by  $A$  and  $B$ . The Wright–Fisher model is based on the sampling of the gene pool at generation  $t$ , which consists of  $n$  genes of type  $A$  and  $(N-n)$  genes of type  $B$ , to produce the next generation of  $N$  genes. Although this may be formulated

as a Markov chain, neither Fisher nor Wright did so; this was first carried out by Malécot [31] in 1944. Here we will describe a variant of the model introduced by Moran [36,37], which is a one-step process and can be formulated as a master equation.

The Moran model does not have non-overlapping generations, as in the Wright–Fisher model, and is more akin to a birth–death process where birth and death are coupled. At a given time, two individuals are sampled with replacement: one is designated the parent which is copied to create an offspring and the other is sacrificed to make way for the new offspring. Clearly if a  $B$  (chosen with probability  $(N-n)/N$ ) is sacrificed and an  $A$  (chosen with probability  $n/N$ ) is copied, this gives a contribution to  $T(n+1|n)$ . If the choice is that with  $A$  and  $B$  interchanged, this gives a contribution to  $T(n-1|n)$ . The transition rates for the Moran model are thus

$$\begin{aligned} T(n+1|n) &= \beta \left(1 - \frac{n}{N}\right) \left(\frac{n}{N}\right), \\ T(n-1|n) &= \beta \left(\frac{n}{N}\right) \left(1 - \frac{n}{N}\right), \end{aligned} \quad (26)$$

where  $\beta$  is a rate constant, which may be absorbed into the time  $t$ .

This may be extended in various ways. For example, mutations may be included:  $A \xrightarrow{u} B$  and  $B \xrightarrow{v} A$ . With probability  $(1-u-v)$  the offspring is taken to be a copy of the parent without mutation, as previously described. For the rest of the time (that is, with probability  $u+v$ ), a mutation occurs. If the parent is an  $A$ , the offspring becomes a  $B$  with probability  $u/(u+v)$ , and if the parent is a  $B$ , the offspring becomes an  $A$  with probability  $v/(u+v)$ . This leads to the transition rates

$$\begin{aligned} T(n+1|n) &= (1-u-v) \left(1 - \frac{n}{N}\right) \left(\frac{n}{N}\right) \\ &\quad + v \left(1 - \frac{n}{N}\right), \\ T(n-1|n) &= (1-u-v) \left(1 - \frac{n}{N}\right) \left(\frac{n}{N}\right) + u \left(\frac{n}{N}\right). \end{aligned} \quad (27)$$

The master equation for the Moran model will be discussed again in the next section.

3. *Competition in a single species model.* The birth–death process described in Example 1 can be generalized to more complex ecological situations. As it stands it consists of the two processes  $A \xrightarrow{d} E$  and  $A \xrightarrow{b} A+A$  representing death and birth respectively. Here  $A$  represents an individual and  $E$  is a null state. To model the finite resources available in a given patch, we

put a limit on the number of allowed individuals:  $n = 0, 1, \dots, N$ . We can also only allow a birth if enough space and/or other resources are available:  $A + E \xrightarrow{b} A + A$  and include competition for these resources:  $A + A \xrightarrow{c} A + E$ . Since the probability of obtaining an  $A$  when sampling the patch is  $n/N$  and of obtaining an  $E$  is  $(N - n)/N$ , the birth term is now proportional to  $n(N - n)/[N(N - 1)]$  and the competition term proportional to  $n(n - 1)/[N(N - 1)]$ . This gives the transition rates to be

$$\begin{aligned} T(n + 1|n) &= \frac{2bn(N - n)}{N(N - 1)}, \\ T(n - 1|n) &= \frac{cn(n - 1)}{N(N - 1)} + \frac{dn}{N}. \end{aligned} \quad (28)$$

This approach can be extended to more than one species, for instance competition between and within two species [34] or predator-prey interactions [35]. In these cases the state space is multi-dimensional:  $\mathbf{n} = (n_1, n_2, \dots)$ . The master equation still has the form (21), but with  $n$  replaced everywhere by the vector  $\mathbf{n}$ .

Whether or not a stationary state of the master equation exists depends on the nature of the boundary conditions. There are many types of boundary conditions, but two are particularly important. If the boundaries are reflecting, then the probability current vanishes there. If they are absorbing, then the probability of being at the boundary is zero. In the former case probability is conserved, in the latter case it is not, and leaks out of the system.

So to find a non-zero pdf as  $t \rightarrow \infty$  (a stationary distribution) we therefore assume that the system lies within two reflecting boundaries. For a one-step process, the net flow of probability from the state  $n$  to the state  $n + 1$ , is  $J(n, t) = g_n P(n, t) - r_{n+1} P(n + 1, t)$ , where  $J(n, t)$  is the probability current. The master Eq. (25) may be written as

$$\frac{dP(n, t)}{dt} = J(n - 1, t) - J(n, t).$$

For a stationary state, the left-hand side of this equation is zero, and the currents will be time-independent. Therefore  $J(n - 1) = J(n)$  for all  $n$ , that is, all the currents are equal. Since the current vanishes at the boundaries, this constant must be zero. Therefore, for reflecting boundary conditions,  $r_{n+1} P_{\text{st}}(n + 1) = g_n P_{\text{st}}(n)$  for all  $n$ . If we suppose that one boundary is at  $n = 0$  and the other at  $n = N$ , then we have that  $P_{\text{st}}(1) = (g_0/r_1)P_{\text{st}}(0)$ ,  $P_{\text{st}}(2) = (g_1/r_2)P_{\text{st}}(1), \dots$ , which implies  $P_{\text{st}}(2) = (g_1 g_0)/(r_2 r_1)P_{\text{st}}(0), \dots$ . Iterating, the stationary

state can be expressed as a simple product:

$$P_{\text{st}}(n) = \frac{g_{n-1} g_{n-2} \dots g_0}{r_n r_{n-1} \dots r_1} P_{\text{st}}(0), \quad n = 1, \dots, N. \quad (29)$$

The constant  $P_{\text{st}}(0)$  is determined by normalization:

$$\begin{aligned} \sum_{n=0}^N P_{\text{st}}(n) &= P_{\text{st}}(0) + \sum_{n>0} P_{\text{st}}(n) = 1 \\ \Rightarrow (P_{\text{st}}(0))^{-1} &= 1 + \sum_{n=1}^N \frac{g_{n-1} g_{n-2} \dots g_0}{r_n r_{n-1} \dots r_1}. \end{aligned} \quad (30)$$

As an example, we return to the Ehrenfest urn (12), which in the language of the master equation is defined by  $g_n = (N - n)/N$  and  $r_n = n/N$  (any overall rate may be absorbed into the time, and this is irrelevant as far as the stationary state is concerned). Here  $n = 0, 1, \dots, N$  and the molecules never go outside this range, so the boundaries are reflecting. Applying Eqs. (29) and (30) shows that the stationary state is the binomial distribution given by Eq. (15).

## The Fokker–Planck Equation

The Fokker–Planck equation describes stochastic processes at a more coarse grained level than those that we have discussed so far. It only involves continuous stochastic variables; these could be for instance the fraction of individuals or genes of a certain kind in a population, whereas the master equation recognized the individuals or genes as discrete entities. To obtain the Fokker–Planck equation we first derive the Kramers–Moyal expansion [29,38].

We begin by defining the *jump moments* for the system:

$$M_\ell(x, t, \Delta t) = \int d\xi (\xi - x)^\ell P(\xi, t + \Delta t|x, t). \quad (31)$$

We will assume that these are known, that is, they can be obtained by some other means. They will, however, only be required in the limit of small  $\Delta t$ .

The starting point for the derivation is the Chapman–Kolmogorov equation (4), with the choice of variables analogous to that used in Eq. (17) for the discrete case:

$$P(x, t + \Delta t) = \int dx' P(x, t + \Delta t|x', t) P(x', t), \quad (32)$$

again dropping the dependence on the initial conditions. The integrand may be written as

$$\begin{aligned} & P(x, t + \Delta t | x', t) P(x', t) \\ &= P([x - \Delta x] + \Delta x, t + \Delta t | [x - \Delta x], t) P([x - \Delta x], t) \\ &= \sum_{\ell=0}^{\infty} \frac{(-1)^\ell}{\ell!} (\Delta x)^\ell \\ &\quad \cdot \frac{\partial^\ell}{\partial x^\ell} \{P(x + \Delta x, t + \Delta t | x, t) P(x, t)\} , \quad (33) \end{aligned}$$

where  $\Delta x = x - x'$ . Integrating over  $x'$  gives for Eq. (32):

$$P(x, t + \Delta t) = \sum_{\ell=0}^{\infty} \frac{(-1)^\ell}{\ell!} \frac{\partial^\ell}{\partial x^\ell} \{M_\ell(x, t, \Delta t) P(x, t)\} . \quad (34)$$

Since  $P(\xi, t | x, t) = \delta(\xi - x)$ , it follows from Eq. (31), that  $\lim_{\Delta t \rightarrow 0} M_\ell(x, t, \Delta t) = 0$  for  $\ell \geq 1$ . Also  $M_0(x, t, \Delta t) = 1$ . Bearing these results in mind, we will now assume that the jump moments for  $\ell \geq 1$  take the form

$$M_\ell(x, t, \Delta t) = D^{(\ell)}(x, t) \Delta t + o(\Delta t) . \quad (35)$$

Substituting this into Eq. (34), dividing by  $\Delta t$  and taking the limit  $\Delta t \rightarrow 0$  gives

$$\frac{\partial P}{\partial t} = \sum_{\ell=1}^{\infty} \frac{(-1)^\ell}{\ell!} \frac{\partial^\ell}{\partial x^\ell} \{D^{(\ell)}(x, t) P(x, t)\} . \quad (36)$$

Equation (36) is the Kramers–Moyal expansion. So far nothing has been assumed other than the Markov property and the existence of Taylor series expansions. However, in many situations, examination of the jump moments reveal that in a suitable approximation they may be neglected for  $\ell > 2$ . In this case, we may truncate the Kramers–Moyal expansion (36) at second order and obtain the *Fokker–Planck equation*:

$$\frac{\partial P}{\partial t} = -\frac{\partial}{\partial x} [A(x, t) P(x, t)] + \frac{1}{2} \frac{\partial^2}{\partial x^2} [B(x, t) P(x, t)] , \quad (37)$$

where  $A = D^{(1)}$  and  $B = D^{(2)}$  are independent of  $t$  if the process is stationary.

To calculate the jump moments (31), it is convenient to write them in terms of the underlying stochastic process  $x(t)$ . We use the notation

$$\langle x(t) \rangle_{x(t_0)=x_0} = \int dx x P(x, t | x_0, t_0) , \quad (38)$$

for the mean of the stochastic variable at time  $t$ , conditional on the value of  $x(t)$  being given to be  $x_0$  at time  $t_0$ . With this notation  $x(t)$  denotes the process and the angle brackets are averages over realizations of this process. More generally, we may define  $\langle f(x(t)) \rangle$  in a similar way, and in particular the jump moments are given by

$$M_\ell(x, t, \Delta t) = \langle (x(t + \Delta t) - x)^\ell \rangle_{x(t)=x} . \quad (39)$$

### Examples of Fokker–Planck Equations

1. *Simple diffusion*. For the simple symmetric random walk,  $g_n = 1$  and  $r_n = 1$  when expressed in the language of the master equation (after a rescaling of the time so that the rates may taken to be equal to unity). From Eqs. (18) and (19) we find that for a one-step stationary process,

$$\begin{aligned} & \langle (n(t + \Delta t) - n)^\ell \rangle_{n(t)=n} \\ &= \begin{cases} (g_n - r_n) \Delta t + o(\Delta t) , & \text{if } \ell \text{ is odd} \\ (g_n + r_n) \Delta t + o(\Delta t) , & \text{if } \ell \text{ is even} , \end{cases} \quad (40) \end{aligned}$$

and so for the symmetric random walk the odd moments all vanish, and the even moments are equal to  $2\Delta t + o(\Delta t)$ . We now make the approximation which will yield the Fokker–Planck equation: we let  $x = nL$ , where  $L$  is the step size, and let  $L \rightarrow 0$ . Since, for  $\ell$  even,  $\langle (x(t + \Delta t) - x)^\ell \rangle = (2L^\ell) \Delta t$ , if we rescale the time by introducing  $\tau = L^2 t$ , then all jump moments higher than the second disappear in the limit  $L \rightarrow 0$ , and Eq. (36) becomes

$$\frac{\partial P}{\partial t} = \frac{\partial^2 P}{\partial x^2} . \quad (41)$$

This is the familiar diffusion equation obtained from a continuum approximation to the discrete random walk.

2. *The diffusion limit of the Moran model*. For the Moran model with no mutation, we have from Eqs. (26) and (40) that the odd moments again vanish. If we describe the process by  $x(t) = n(t)/N$ , the fraction of the genes that are of type  $A$  at time  $t$ , then the even jump moments are given by

$$\begin{aligned} & \langle (x(t + \Delta t) - x)^\ell \rangle_{x(t)=x} \\ &= \frac{1}{N^\ell} 2x(1-x) \Delta t + o(\Delta t) , \end{aligned}$$

and so introducing a rescaled time  $\tau = 2t/N^2$ , and letting  $N \rightarrow \infty$  we obtain the Fokker–Planck equation

$$\frac{\partial P}{\partial t} = \frac{1}{2} \frac{\partial^2}{\partial x^2} [x(1-x)P] . \quad (42)$$

The factor of 2 in the rescaling of the time is included so that the diffusive form of the Moran model agrees with that found in the Wright–Fisher model [6].

Now suppose mutations are included. The transition rates are given by Eq. (27) and lead to jump moments  $\langle (x(t + \Delta t) - x)^\ell \rangle \sim N^{-\ell}$ . So the first and second jump moments are not of the same order, and the introduction of a rescaled time  $\tau = t/N$ , and subsequently letting  $N \rightarrow \infty$ , gives a Fokker–Planck equation of the form (37), but with  $B = 0$  and  $A(x) = v - (u + v)x$ . This corresponds to a *deterministic* process with  $dx/dt = v - (u + v)x$  [22]. The fact that the system tends to a macroscopic equation when  $N \rightarrow \infty$  has to be taken into account when determining the nature of the fluctuations for large  $N$ . We will discuss this further in Sect. “System Size Expansion”.

On the other hand, suppose that the mutation rates scale with  $N$  according to  $u = 2\tilde{u}/N$  and  $v = 2\tilde{v}/N$ , where  $\tilde{u}$  and  $\tilde{v}$  have a finite limit as  $N \rightarrow \infty$ , and where the 2 has again been chosen to agree with the Wright–Fisher model. Now both the first and second jump moments are of order  $N^{-2}$ , with the higher moments falling off faster with  $N$ . Therefore once again introducing the rescaled time  $\tau = 2t/N^2$  and letting  $N \rightarrow \infty$ , we obtain the Fokker–Planck equation

$$\frac{\partial P}{\partial t} = -\frac{\partial}{\partial x} [\{\tilde{v} - (\tilde{u} + \tilde{v})x\}P] + \frac{1}{2} \frac{\partial^2}{\partial x^2} [x(1-x)P]. \quad (43)$$

So depending on the precise scaling with  $N$ , the Moran model gives different limits as  $N \rightarrow \infty$  [24]. In the first case, the mutational effects are strong enough that a macroscopic description exists, with fluctuations about the macroscopic state, as discussed in Sect. “System Size Expansion”. In the second case, the mutational effects are weaker, and there is no macroscopic equation describing the system; the large  $N$  limit is a nonlinear Fokker–Planck equation of the diffusive type.

The Fokker–Planck equation (37) for stationary processes, where  $A$  and  $B$  are functions of  $x$  only, can be solved by separation of variables, with solutions of the form  $p(x)e^{-\lambda t}$ , where  $\lambda$  is a constant. The equation for  $p(x)$  is then a second order differential equation which, when boundary conditions are given, is of the Sturm–Liouville type. To specify the boundary conditions we once again introduce the probability current, this time through the continuity equation,

$$\frac{\partial P(x, t)}{\partial t} + \frac{\partial J(x, t)}{\partial x} = 0, \quad (44)$$

where the probability current  $J(x, t)$  is given by

$$J(x, t) = A(x, t)P(x, t) - \frac{1}{2} \frac{\partial}{\partial x} [B(x, t)P(x, t)]. \quad (45)$$

Let us suppose that the system is defined on the interval  $[a, b]$ . Then if the boundaries are reflecting, there is no net flow of probability across  $x = a$  and  $x = b$ . This implies that  $J(a, t) = 0$  and  $J(b, t) = 0$ . If we integrate the equation of continuity (44) from  $x = a$  to  $x = b$ , and apply these boundary conditions, we see that  $\int_a^b P(x, t) dx$  is independent of time. Therefore, if the pdf is initially normalized, it remains normalized. This is in contrast with the case of absorbing boundary conditions, defined by  $P(a, t) = 0$  and  $P(b, t) = 0$ . If the boundary conditions are at infinity we require that  $\lim_{x \rightarrow \pm\infty} P(x, t) = 0$ , so that if  $P$  is well-behaved it is normalizable, and also that  $\partial P/\partial x$  is well-behaved in this limit:  $\lim_{x \rightarrow \pm\infty} \partial P/\partial x = 0$ . If  $A$  or  $B$  do not diverge as  $x \rightarrow \pm\infty$  this implies that  $\lim_{x \rightarrow \pm\infty} J(x, t) = 0$ . Other types of boundary conditions are possible, and we do not attempt a complete classification here [16, 46].

If  $A$  and  $B$  are independent of time, then from Eq. (44), the stationary state of the system must be given by  $dJ(x)/dx = 0$ , that is,  $J$  is a constant. For reflecting boundary conditions this constant is zero, and so from Eq. (45) the stationary pdf,  $P_{\text{st}}(x)$  must satisfy

$$0 = A(x)P_{\text{st}}(x) - \frac{1}{2} \frac{\partial}{\partial x} [B(x)P_{\text{st}}(x)]. \quad (46)$$

This may be integrated to give

$$P_{\text{st}}(x) = \frac{C}{B(x)} \exp \left\{ 2 \int^x dx' \frac{A(x')}{B(x')} \right\}, \quad (47)$$

where  $C$  is a constant which has to be chosen so that  $P_{\text{st}}(x)$  is normalized.

The Fokker–Planck equation, with  $A$  independent of  $t$  and  $B$  constant, can be transformed into the Schrödinger-like problem

$$-B \frac{\partial \psi}{\partial t} = -\frac{B^2}{2} \frac{\partial^2 \psi}{\partial x^2} + U(x)\psi, \quad (48)$$

by the transformation

$$P(x, t) = [P_{\text{st}}(x)]^{1/2} \psi(x, t), \quad (49)$$

where

$$U(x) = \frac{1}{2} [A(x)]^2 + \frac{B}{2} \frac{dA}{dx}. \quad (50)$$

So a one-dimensional stationary stochastic process, under certain conditions (such as constant second jump moment) is equivalent to quantum mechanics in imaginary time, with  $B$  taking over the role of Planck’s constant.

As an example we consider the *Ornstein–Uhlenbeck process* defined by

$$\frac{\partial P}{\partial t} = \frac{\partial}{\partial x} [axP] + D \frac{\partial^2 P}{\partial x^2}, \quad x \in (-\infty, \infty), \quad a > 0. \quad (51)$$

In this case, the potential (50) is given by  $U(x) = [a^2x^2 - 2aD]/2$ , and so the problem is equivalent to the one-dimensional simple harmonic oscillator in quantum mechanics, but with an energy shift. As in that problem [47], the eigenfunctions are Hermite polynomials. Specifically, the right-eigenfunctions are

$$p_m(x) = P_{\text{st}}(x) \frac{1}{[2^m m!]^{1/2}} H_m(\alpha x), \quad (52)$$

where  $H_m$  are the Hermite polynomials [1] and  $\alpha = (a/2D)^{1/2}$ . The eigenvalue corresponding to the eigenfunction (52) is  $\lambda_m = am$ ,  $m = 0, 1, \dots$  and the left-eigenfunctions are  $q_m(x) = [P_{\text{st}}(x)]^{-1} p_m(x)$ . From these explicit solutions we may calculate other quantities of interest, such as correlation functions. We note that the eigenvalues are all non-negative and that the stationary state corresponds to  $\lambda = 0$ . The left-eigenfunction for the stationary state is equal to 1. These latter results hold true for a wide-class of such problems.

### Stochastic Differential Equations

So far we have described stochastic processes in terms of equations which give the time evolution of pdfs. In this section, we will describe equations for the stochastic variables themselves. The most well known instance of such an equation is the Langevin equation for the velocity of a Brownian particle, and so we begin with this particular example.

Suppose that a small macroscopic particle of mass  $m$  (such as a pollen grain) is immersed in a liquid at a temperature  $T$ . In addition to any macroscopic motion that the particle may have, its velocity fluctuates due to the random collisions of the particle with the molecules of the liquid. For simplicity, we confine ourselves to one-dimensional motion – along the  $x$ -axis. Then the equation of motion of the particle may be written in the form

$$m \frac{d^2x}{dt^2} = -\alpha \frac{dx}{dt} - \frac{dV}{dx} + \mathcal{F}(t). \quad (53)$$

The first term on the right-hand side is due to the viscosity of the fluid and  $\alpha$  is the friction constant. The second term, where  $V(x)$  is a potential, represents the interaction of the particle with any external forces, such as gravity. The final term is the random force due to collisions with the

molecules of the liquid. Clearly to complete the specification of the dynamics of the particle we need to give (i) the initial position and velocity of the particle, and (ii) the statistics of the random force  $\mathcal{F}(t)$ .

To make progress with these points, we imagine a large number of realizations of the dynamics, in which the particle starts with the same initial position,  $x_0$ , and velocity,  $v_0$ , but where the initial positions and velocities of the molecules in the liquid will be different. Taking the average over a large number of such realizations will give the average position  $\langle x(t) \rangle$  and velocity  $\langle v(t) \rangle$  at time  $t$ , conditional on  $x(0) = x_0$  and  $v(0) = v_0$ . The statistics of the fluctuating force  $\mathcal{F}(t)$  are assumed to be such that

- (a)  $\langle \mathcal{F}(t) \rangle = 0$ , since we do not expect one direction to be favored over the other.
- (b)  $\langle \mathcal{F}(t)\mathcal{F}(t') \rangle = 2D\delta(t - t')$ , since we expect that after a few molecular collisions the value that  $\mathcal{F}$  takes on will be independent of its former value. That is, the force  $\mathcal{F}$  becomes uncorrelated over times of the order of a few collision times between molecules. This is tiny on observational time scales, and so taking the correlation function to be a delta-function is an excellent approximation. The weight of the delta-function is denoted by  $2D$ , where at this stage  $D$  is undetermined.
- (c)  $\mathcal{F}(t)$  is taken to be Gaussianly distributed on grounds of simplicity, but also because by the central limit theorem it is assumed that the net effect of the large number of molecules which collide with the pollen grain will lead to a distribution which is Gaussian.

Since a Gaussian distribution is specified by its first two moments, conditions (a), (b) and (c) completely define the statistics of  $\mathcal{F}(t)$ .

Finally, Eq. (53) as it stands does not define a Markov process. This is most easily seen if we write down a discrete time version of the equation. The second derivative means that  $x(t + \delta t)$  not only depends on  $x(t)$ , but also on  $x(t - \delta t)$ . Therefore only first order derivatives should be included in such equations if the process is to be Markov. This is easily achieved by promoting  $v(t)$  to be a second stochastic variable in addition to  $x(t)$ . Then Eq. (53) may be equivalently written as

$$\begin{aligned} \frac{dx}{dt} &= v, \\ m \frac{dv}{dt} &= -\alpha v - \frac{dV}{dx} + \mathcal{F}(t), \end{aligned} \quad (54)$$

which does define a Markov process. Although, as we remarked in the Introduction, we deal almost exclusively with Markov processes in this article, the situation we have just discussed is a good illustration of one way of



dealing with processes which are presented as being non-Markovian. The method simply consists of adding a sufficient number of supplementary variables to the definition of the state variables of the process until it becomes Markovian. There is no guarantee that this will be possible or require only a small number of additional variables to be promoted in this way, but it is the most straightforward and direct way of rendering non-Markovian processes tractable.

To begin the analysis of Eq. (54) we assume that there are no external forces and so the term  $dV/dx$  is equal to zero. We may then write Eq. (54) as the Langevin equation

$$\frac{dv}{dt} = -\gamma v + F(t); \quad v(0) = v_0, \quad (55)$$

where  $\gamma = \alpha/m$  and  $F(t) = \mathcal{F}(t)/m$ . This implies that

$$\langle F(t) \rangle = 0 \quad \text{and} \quad \langle F(t)F(t') \rangle = \frac{2D}{m^2} \delta(t - t'). \quad (56)$$

Note that since  $F(t)$  is a random variable, solving the Langevin equation will give  $v(t)$  as a random variable (having a known distribution). It is therefore a *stochastic differential equation*. The function  $F$  is frequently called “the noise term” or simply “the noise”. It is *white noise* since the Fourier transform of a delta-function is a constant – all frequencies are present in equal amounts.

Multiplying the Langevin equation (55) by the integrating factor  $e^{\gamma t}$  gives

$$\begin{aligned} \frac{d}{dt} [v(t)e^{\gamma t}] &= F(t)e^{\gamma t} \Rightarrow v(t) \\ &= v_0 e^{-\gamma t} + e^{-\gamma t} \int_0^t dt' F(t') e^{\gamma t'}. \end{aligned} \quad (57)$$

By taking the average of the expression for  $v(t)$  we find  $\langle v(t) \rangle = v_0 e^{-\gamma t}$ . More interestingly, if we square the expression for  $v(t)$  and take the average, then we find

$$\langle v^2(t) \rangle = v_0^2 e^{-2\gamma t} + \frac{D}{\alpha m} [1 - e^{-2\gamma t}], \quad (58)$$

which implies that

$$\lim_{t \rightarrow \infty} \langle v^2(t) \rangle = \frac{D}{\alpha m}. \quad (59)$$

On the other hand, as  $t \rightarrow \infty$ , the Brownian particle will be in thermal equilibrium:

$$\lim_{t \rightarrow \infty} \langle v^2(t) \rangle = v_{\text{eq}}^2 \quad \text{and} \quad \frac{1}{2} m v_{\text{eq}}^2 = \frac{1}{2} kT,$$

where  $T$  is the temperature of the liquid and  $k$  is Boltzmann's constant. This implies that

$$\frac{1}{2} m \left( \frac{D}{\alpha m} \right) = \frac{1}{2} kT \Rightarrow D = \alpha kT. \quad (60)$$

The molecules of the liquid are acting as a heat bath for the system – which in this case is a single Brownian particle. The equation  $D = \alpha kT$  is a simple example of a fluctuation-dissipation theorem, and determines  $D$  in terms of the friction constant,  $\alpha$ , and of the temperature of the liquid,  $T$ .

Although we have presented a somewhat heuristic rationale for Eq. (54), it may be derived in a more controlled way. A particularly clear derivation has been given by Zwanzig [60], where the starting point is a Hamiltonian which contains three terms: for the system, the heat bath and the interaction between the system and the heat bath. Taking the heat bath to be made up of coupled harmonic oscillators and the interaction term between the system and heat bath to be linear, it is possible to integrate out the bath degrees of freedom exactly, and be left only with the equations of motion of the system degrees of freedom plus the initial conditions of the bath degrees of freedom. Assuming that the bath is initially in thermal equilibrium, so that these initial values are distributed according to a Boltzmann distribution, adds extra “noise” terms to the equations of motion which, with a few more plausible assumptions, make them of the Langevin type.

### Examples of Langevin-like Equations

1. *Overdamped Brownian motion.* Frequently the viscous damping force  $-\alpha v$  is much larger than the inertial term  $m d^2x/dt^2$  in Eq. (53), and so to a good approximation the left-hand side of Eq. (53) can be neglected. Scaling time by  $\alpha$ , we arrive at the Langevin equation for the motion of an overdamped Brownian particle:

$$\frac{dx}{dt} = -V'(x) + \mathcal{F}(t); \quad x(0) = x_0, \quad (61)$$

where  $'$  denotes differentiation with respect to  $x$  and where, due to the rescaling of time by  $\alpha$ ,

$$\langle \mathcal{F}(t) \rangle = 0; \quad \langle \mathcal{F}(t)\mathcal{F}(t') \rangle = 2\tilde{D}\delta(t - t'); \quad \tilde{D} = \frac{D}{\alpha}. \quad (62)$$

A particularly well-known case is when the Brownian particle is moving in the harmonic potential  $V(x) = ax^2/2$ . Then

$$\frac{dx}{dt} = -ax + \mathcal{F}(t); \quad x(0) = x_0. \quad (63)$$

Since Eq. (63) relating  $x(t)$  to  $\mathcal{F}(t)$  is linear, and since the distribution of  $\mathcal{F}(t)$  is Gaussian, then  $x(t)$  is also distributed according to a Gaussian distribution. Comparing with Eqs. (55) and (56), which also define a linear system, we find that  $\langle x(t) \rangle = x_0 e^{-at}$  and, from

Eq. (58),  $\langle x^2(t) \rangle = x_0^2 e^{-2at} + (\tilde{D}/a) [1 - e^{-2at}]$ . This gives

$$P(x, t | x_0, 0) = \sqrt{\frac{a}{2\pi\tilde{D}[1 - e^{-2at}]}} \exp \left\{ -\frac{a(x - x_0 e^{-at})^2}{2\tilde{D}[1 - e^{-2at}]} \right\}. \quad (64)$$

It is straightforward to check that this conditional pdf satisfies the Fokker–Planck equation (51) for the Ornstein–Uhlenbeck process. Below we will show this more directly, by starting from the Langevin equation (61) and deriving Eq. (51) if  $V(x)$  is quadratic.

Another case of interest is when  $V(x)$  is a double-well potential,  $V(x) = -ax^2/2 + bx^4/4$ . If the particle is initially located near the bottom of one of the potential wells, it will take on average a time of the order of  $e^{\Delta V/D}$  to hop over the barrier and into the well on the other side. Here  $\Delta V$  is the height of the barrier that it has to hop over [29].

2. *Environmental noise in population biology.* One of the simplest models of two species with population sizes  $N_1$  and  $N_2$  which are competing for a common resource, is the two coupled deterministic ordinary differential equations  $\dot{N}_i = r_i N_i$ ,  $i = 1, 2$ . The growth rates,  $r_i$ , depend on the population sizes in such a way that as the population sizes increase, the  $r_i$  decrease to reflect the increased competition for resources. This could be modeled, for instance, by taking  $r_i = a_i - b_{ii}N_i - b_{ij}N_j$  with  $i, j = 1, 2$  and  $i \neq j$ . In reality, external factors such as climate, terrain, the presence of other species, and indeed any factor which has an uncertain influence on these two species, will also affect the growth rate. This can be modeled by adding an external random term to the  $r_i$  which represents this *environmental stochasticity* [33]. Then the equations become

$$\begin{aligned} \frac{dN_1}{dt} &= a_1 N_1 - b_{11} N_1^2 - b_{12} N_1 N_2 + N_1 \zeta_1(t) \\ \frac{dN_2}{dt} &= a_2 N_2 - b_{22} N_2^2 - b_{21} N_2 N_1 + N_2 \zeta_2(t). \end{aligned} \quad (65)$$

Since the noise terms,  $\zeta_i(t)$  are designed to reflect the large number of coupled variables omitted from the description of the model, it is natural, by virtue of the central limit theorem, to assume that they are Gaussianly distributed. It also seems reasonable to assume that any temporal correlation between these external influences is on scales very much shorter than those of interest to

us here, and that the noises have zero mean. We therefore assume that

$$\langle \zeta_i(t) \rangle = 0; \quad \langle \zeta_i(t) \zeta_j(t') \rangle = 2D_i \delta_{ij} \delta(t - t'), \quad (66)$$

where the  $D_i$  describe the strength of the stochastic effects. The deterministic equations (that is, Eq. (65) without the noise terms) have a fixed point at the origin, one on each of the  $N_1$  and  $N_2$  axes, and may have another at non-zero  $N_1$  and  $N_2$ . For some values of the parameters this latter fixed point may be a saddle, with those on the axes being stable and the origin unstable. In this situation the eventual fate of the species depends significantly on the noise: if the combination of the nonlinear dynamics and the noise drives the system to the vicinity of the fixed point on the  $N_1$  axis, then species 2 will become extinct, and vice-versa.

Langevin equations with Gaussian white noise are equivalent to Fokker–Planck equations. This can be most easily seen by calculating the jump moments (39) from the Langevin equation. For instance, if we begin from the Langevin equation for an overdamped Brownian particle (61),

$$\begin{aligned} \Delta x(t) &\equiv x(t + \Delta t) - x(t) = \int_t^{t+\Delta t} dt' \dot{x}(t') \\ &= - \int_t^{t+\Delta t} dt' V'(x(t')) + \eta(t), \end{aligned} \quad (67)$$

where  $\eta(t) = \int_t^{t+\Delta t} dt' \mathcal{F}(t')$ . From Eq. (62) it is straightforward to calculate the moments of  $\eta(t)$ :  $\langle \eta(t) \rangle = 0$ ,

$$\langle \eta^2(t) \rangle = \int_t^{t+\Delta t} dt' \int_t^{t+\Delta t} dt'' \langle \mathcal{F}(t') \mathcal{F}(t'') \rangle = 2\tilde{D}\Delta t \quad (68)$$

and, since  $\eta(t)$  is Gaussian,  $\langle \eta^n(t) \rangle$  is zero if  $n$  is odd, and at least of order  $(\Delta t)^2$  for  $n \geq 4$ . This implies that

$$\begin{aligned} M_1(x, \Delta t) &= -V'(x)\Delta t + \mathcal{O}(\Delta t)^2, \\ M_2(x, \Delta t) &= 2\tilde{D}\Delta t + \mathcal{O}(\Delta t)^2, \end{aligned} \quad (69)$$

with all moments of order  $(\Delta t)^2$  or higher for  $\ell > 2$ . The notation  $\mathcal{O}(\Delta t)^2$  means that the magnitude of this quantity is less than a constant times  $(\Delta t)^2$ , for sufficiently small nonzero  $(\Delta t)^2$ . This is a weaker, but more specific, statement than saying it is  $o(\Delta t)$ . Using Eqs. (35) and (36), the Fokker–Planck equation which is equivalent to the

Langevin equation (61) is found to be

$$\frac{\partial P}{\partial t} = \frac{\partial}{\partial x} [V'(x)P] + \tilde{D} \frac{\partial^2 P}{\partial x^2}. \quad (70)$$

From Eq. (47), the stationary pdf is  $P_{\text{st}}(x) = C \exp\{-V(x)/\tilde{D}\} = C \exp\{-V(x)/kT\}$ , as expected.

Although in this article we have largely restricted our attention to stochastic processes involving one variable, the construction of a Fokker–Planck equation from the Langevin equation goes through in a similar way for an  $n$ -dimensional process  $\mathbf{x} = (x_1, \dots, x_n)$ . In this case the jump moments are

$$\begin{aligned} \langle \Delta x_{i_1}(t) \Delta x_{i_2}(t) \dots \Delta x_{i_\ell}(t) \rangle_{\mathbf{x}(t)=\mathbf{x}} \\ = D_{i_1 \dots i_\ell}(\mathbf{x}, t) \Delta t + o(\Delta t), \end{aligned} \quad (71)$$

where  $\Delta x_{i_\alpha} = x_{i_\alpha}(t + \Delta t) - x_{i_\alpha}$ . The Kramers–Moyal expansion is then

$$\frac{\partial P}{\partial t} = \sum_{\ell=1}^{\infty} \frac{(-1)^\ell}{\ell!} \frac{\partial^\ell}{\partial x_{i_1} \dots \partial x_{i_\ell}} \{D_{i_1 \dots i_\ell}(\mathbf{x}, t) P\}. \quad (72)$$

The Langevin equation for Brownian motion (54), without going to the overdamped limit, serves as a simple illustration of this generalization. Here  $\Delta x(t) = v \Delta t$  and  $\Delta v(t) = -\gamma v \Delta t - m^{-1} V'(x) \Delta t + m^{-1} \eta(t)$ . This results in the Fokker–Planck equation

$$\frac{\partial P}{\partial t} = -\frac{\partial}{\partial x} [vP] + \frac{\partial}{\partial v} [\{\gamma v + m^{-1} V'(x)\} P] + \frac{\gamma kT}{m} \frac{\partial^2 P}{\partial v^2}. \quad (73)$$

This is Kramer’s equation. It has a stationary pdf  $P_{\text{st}}(x, v) = C \exp\{-E/kT\}$ , where  $E = mv^2/2 + V(x)$ .

We end this section by finding the Fokker–Planck equation which is equivalent to the general set of Langevin equations of the form

$$\dot{x}_i = A_i(\mathbf{x}, t) + \sum_{\alpha=1}^m g_{i\alpha}(\mathbf{x}, t) \zeta_\alpha(t); \quad i = 1, \dots, n, \quad (74)$$

where  $\zeta_\alpha(t)$ ,  $\alpha = 1, \dots, m$ , is a Gaussian white noise with zero mean and with

$$\langle \zeta_\alpha(t) \zeta_\beta(t') \rangle = \delta_{\alpha\beta} \delta(t - t'). \quad (75)$$

Proceeding as in Eq. (67), but noting the dependence of the function  $g_{i\alpha}$  on the stochastic variable, yields

$$\begin{aligned} M_i(\mathbf{x}, t, \Delta t) \\ = \left[ A_i(\mathbf{x}, t) + \theta(0) \sum_{j=1}^n \sum_{\alpha=1}^m g_{j\alpha}(\mathbf{x}, t) \frac{\partial}{\partial x_j} g_{i\alpha}(\mathbf{x}, t) \right] \Delta t \\ + \mathcal{O}(\Delta t)^2, \\ M_{ij}(\mathbf{x}, t, \Delta t) = \sum_{\alpha=1}^m [g_{i\alpha}(\mathbf{x}, t) g_{j\alpha}(\mathbf{x}, t)] \Delta t + \mathcal{O}(\Delta t)^2, \end{aligned} \quad (76)$$

with all jump moments higher than the second being of order  $(\Delta t)^2$  or higher. The quantity  $\theta(0)$  is the value of the Heaviside theta function,  $\theta(x)$ , at  $x = 0$  and is indeterminate. This indicates that the Langevin description does not correspond to a unique Fokker–Planck equation. This situation occurs whenever the white noise in a Langevin equation is multiplied by a function which depends on the state variable, as in Eq. (74). For systems such as this acted upon by *multiplicative noise* the Langevin description has to be supplemented by a rule which says whether the state variable in the multiplying function ( $g_{i\alpha}$  in Eq. (74)) is that before or after the noise pulse acts [52]. If it is taken to be the value immediately after the noise pulse acts then  $\theta(0) = 0$  (Itô rule), whereas if it is taken to be the average of the values before and after, then  $\theta(0) = 1/2$  (Stratonovich rule). The Fokker–Planck equation is now found from Eq. (72) to be

$$\begin{aligned} \frac{\partial P}{\partial t} = - \sum_{i=1}^n \frac{\partial}{\partial x_i} [A_i(\mathbf{x}, t) P(\mathbf{x}, t)] \\ + \frac{1}{2} \sum_{i,j=1}^n \sum_{\alpha=1}^m \frac{\partial^2}{\partial x_i \partial x_j} [g_{i\alpha}(\mathbf{x}, t) g_{j\alpha}(\mathbf{x}, t) P(\mathbf{x}, t)], \end{aligned} \quad (77)$$

in the Itô case and

$$\begin{aligned} \frac{\partial P}{\partial t} = - \sum_{i=1}^n \frac{\partial}{\partial x_i} [A_i(\mathbf{x}, t) P(\mathbf{x}, t)] \\ + \frac{1}{2} \sum_{i,j=1}^n \sum_{\alpha=1}^m \frac{\partial}{\partial x_i} \left[ g_{i\alpha}(\mathbf{x}, t) \frac{\partial}{\partial x_j} \{g_{j\alpha}(\mathbf{x}, t) P(\mathbf{x}, t)\} \right], \end{aligned} \quad (78)$$

in the Stratonovich case.

### Path Integrals

While most early work on stochastic processes was concerned with linear systems, naturally attention soon

moved on to the many interesting systems which could be modeled as nonlinear stochastic processes. These systems are much more difficult to analyze. For example, a nonlinear Langevin equation cannot be solved directly, and so the averaging procedure cannot be carried out in the same explicit way as described in Sect. “Stochastic Differential Equations”. There is however one method which is applicable to many nonlinear stochastic differential equations of interest: the solution of these equations can be formally written down as a path-integral, and from this correlation functions and other quantities of physical interest can be obtained. This also has the advantage that all the formalism and approximation schemes developed to study functional integrals over the years can be called into play.

Path-integrals are intimately related to Brownian motion and the earliest work on the subject by Wiener [56, 57], emphasized this. If the problem of interest is formulated as a set of Langevin equations, the derivation of the path-integral representation is particularly straightforward, if rather heuristic. For clarity we begin with the simplest case: an overdamped system with a single degree of freedom,  $x$ , acted upon by white noise. The Langevin equation is given by Eq. (61) and the noise is defined by Eq. (62). Since the noise is assumed to be Gaussian, Eq. (62) is a complete specification. An equivalent way of giving it is through the pdf [12]:

$$P[\mathcal{F}] \mathcal{D}\mathcal{F} \propto \exp\left(-\frac{1}{4\bar{D}} \int dt \mathcal{F}^2(t)\right) \mathcal{D}\mathcal{F}, \quad (79)$$

where  $\mathcal{D}\mathcal{F}$  is the functional measure. The idea is now to regard the Langevin equation (61) as defining a mapping  $\mathcal{F} \mapsto x$ . The pdf for the  $x$  variable is then given by

$$P[x] = P[\mathcal{F}]|_{\mathcal{F}=\dot{x}+V'(x)} J[x] \propto \exp\left(-\frac{1}{4\bar{D}} \int dt [\dot{x} + V'(x)]^2\right) J[x], \quad (80)$$

where

$$J[x] = \det\left[\frac{\delta \mathcal{F}}{\delta x}\right], \quad (81)$$

is the Jacobian of the transformation. An explicit expression for the Jacobian may be obtained either by direct calculation of a discretized form of the Langevin equation [21] or through use of the identity relating the determinant of a matrix to the exponential of the trace of the logarithm of that matrix [59]. One finds that  $J[x] \propto \exp\{\theta(0) \int dt V''(x)\}$ . The quantity  $\theta(0)$  is once again the indeterminate value of the Heaviside theta function  $\theta(x)$  at  $x = 0$ . Its appearance is a reflection of the fact that,

due to the Brownian-like nature of the paths in the functional integral, the nature of the discretization appears explicitly through this factor [48]. If we consistently use the mid-point rule throughout, then we may take  $\theta(0) = 1/2$ , which gives

$$P[x] \propto \exp\left(-\frac{1}{4\bar{D}} \int dt [\dot{x} + V'(x)]^2 + \frac{1}{2} \int dt V''(x)\right) = \exp(-S[x]/\bar{D}). \quad (82)$$

All quantities of interest can now be found from expression (82). For example, the conditional probability distribution,  $P(x, t|x_0, t_0)$  is given by

$$\langle \delta(x-x(t)) \rangle_{x(t_0)=x_0} = \int_{x(t_0)=x_0} \mathcal{D}x \delta(x-x(t)) P[x]. \quad (83)$$

The expression (82) has much in common with Feynman’s formulation of quantum mechanics as a path-integral [11]. In fact another way to obtain the result is to exploit the transformation (49) to write the Fokker–Planck equation (70) as a Schrödinger equation in imaginary time  $\tau = it$ , with a potential  $U(x) = (1/2)[V'(x)]^2 - \bar{D}V''(x)$ , following Eq. (50). The action in the quantum-mechanical path-integral is

$$\frac{i}{\hbar} \int dt \left[ \frac{1}{2} \dot{x}^2 - U(x) \right] \longrightarrow \frac{1}{2\bar{D}} \int d\tau \left[ -\frac{1}{2} \dot{x}^2 - \frac{1}{2} [V'(x)]^2 + \bar{D}V''(x) \right], \quad (84)$$

which is Eq. (82) since  $\int_{t_0}^t dt \dot{x} V'(x) = \int_{x_0}^x dx V'(x) = V(x) - V(x_0)$  does not depend on the path, only on the end-points. The functional  $S[x]$  is analogous to the action in classical mechanics, and is frequently referred to as such. It is also sometimes referred to as the generalized Onsager–Machlup functional, in recognition of the original work carried out by Onsager and Machlup [40], in the case of a linear Langevin equation, in 1953.

The above discussion can be generalized in many ways. For example, if the Langevin equation for an  $n$ -dimensional process takes the form

$$\dot{x}_i = A_i(\mathbf{x}) + \zeta_i(t), \quad \langle \zeta_i(t) \zeta_j(t') \rangle = 2D_{ij} \delta(t-t'), \quad (85)$$

where  $\zeta_i(t)$  is a Gaussian noise with zero mean and  $D_{ij}$  is independent of  $\mathbf{x}$ , then the general Onsager–Machlup

functional is [21]

$$S[x] = \int dt \left[ \frac{1}{4} \sum_{i,j} \{ \dot{x}_i - A_i(\mathbf{x}) \} D_{ij}^{-1} \{ \dot{x}_j - A_j(\mathbf{x}) \} + \frac{1}{2} \sum_i \frac{\partial A_i}{\partial x_i} \right], \quad (86)$$

if the matrix  $D_{ij}$  is non-singular. The generalization to the situation where the noise is multiplicative is more complicated, and is analogous to the path-integral formulation of quantum mechanics in curved space [20].

### System Size Expansion

In Example 2 of Sect. “The Fokker–Planck Equation” we explicitly showed how the master equation may have different limits when the size of the system,  $N$ , becomes large. In one case both the first and second jump moments were of the same order (and much larger than the higher jump moments) and so a nonlinear Fokker–Planck equation of the diffusion type was obtained in the limit  $N \rightarrow \infty$ . In another case, the first jump moment scaled in a different way to the second moment, and so the  $N \rightarrow \infty$  limit gave a deterministic macroscopic equation of the form  $\dot{x} = f(x)$ , with finite  $N$  effects presumably consisting of  $1/\sqrt{N}$  fluctuations about the macroscopic state,  $x(t)$ . It is the second scenario that we will explore in this section. It can be formalized by writing

$$\frac{n}{N} = x(t) + \frac{\xi}{\sqrt{N}}, \quad (87)$$

and substituting this into the master equation, then equating terms of the same order in  $1/\sqrt{N}$ . The leading order equation obtained in this way will be the macroscopic equation, and the function  $f(x)$  will emerge from the analysis. The next-to-leading order equation turns out to be a linear Fokker–Planck equation in the variable  $\xi$ . Higher order terms may also be included. This formalism was first developed by van Kampen [51] and is usually referred to as van Kampen’s system-size expansion. We will describe it in the specific case of a one-step process for a single stochastic variable in order to bring out the essential features of the method.

When using this formalism it is useful to rewrite the master equation (23) using step operators which act on an arbitrary function of  $n$  according to  $\mathcal{E}f(n) = f(n+1)$  and  $\mathcal{E}^{-1}f(n) = f(n-1)$ . This gives

$$\frac{dP(n, t)}{dt} = (\mathcal{E} - 1) [T(n-1|n)P(n, t)] + (\mathcal{E}^{-1} - 1) [T(n+1|n)P(n, t)]. \quad (88)$$

We begin by using Eq. (87) to write the pdf which appears in the master Equation (88) as

$$P(Nx(t) + \sqrt{N}\xi, t) = \Pi(\xi, t) \\ \Rightarrow \dot{P} = \frac{\partial \Pi}{\partial t} - N^{1/2} \frac{dx}{dt} \frac{\partial \Pi}{\partial \xi}. \quad (89)$$

This gives an expression for the left-hand side of the master equation,  $\dot{P}$ . To get an expression for the right-hand side,

- (a) the step operators are expanded in powers of  $1/\sqrt{N}$  [53]:

$$\mathcal{E}^{\pm 1} = 1 \pm \frac{1}{\sqrt{N}} \frac{\partial}{\partial \xi} + \frac{1}{2!} \frac{1}{N} \frac{\partial^2}{\partial \xi^2} + \mathcal{O}\left(\frac{1}{N^{3/2}}\right), \quad (90)$$

- (b)  $T(n \pm 1|n)$  is expressed in terms of  $\xi$  and  $N$ ,  
(c)  $P(n, t)$  is replaced by  $\Pi(\xi, t)$ .

Steps (a), (b) and (c) gives the right-hand side of the master equation as a power-series in  $1/\sqrt{N}$ . Equating the left-hand and right-hand sides order by order in  $1/\sqrt{N}$  (this may require a rescaling of the time,  $t$ , by a power of  $\sqrt{N}$ ), gives to leading order (the  $\partial \Pi / \partial \xi$  cancels) an equation of the form  $dx/dt = f(x)$ . This may be solved subject to the condition  $x(0) = x_0 = n_0/N$ , if we take the initial condition on the master equation to be  $P(n, 0) = \delta_{n, n_0}$ . We denote the solution of this macroscopic equation by  $x_M(t)$ .

To next order in  $1/\sqrt{N}$ , the Fokker–Planck equation

$$\frac{\partial \Pi}{\partial t} = -f'(x) \frac{\partial}{\partial \xi} [\xi \Pi] + \frac{1}{2} g(x) \frac{\partial^2 \Pi}{\partial \xi^2}, \quad (91)$$

describing a linear stochastic process is found. The functions  $f'(x)$  and  $g(x)$  are to be evaluated when  $x = x_M(t)$ , and so are simply functions of time. If the macroscopic system tends to a fixed point:  $x_M(t) \rightarrow x^*$ , as  $t \rightarrow \infty$ , then  $f'(x)$  and  $g(x)$  may be replaced by constants in order to study the fluctuations about this stationary state.

To illustrate the method we use Example 3, Sect. “The Master Equation”. Equating both sides of the master equation in this case one finds

$$-N^{1/2} \frac{dx}{dt} \frac{\partial \Pi}{\partial \xi} + \frac{\partial \Pi}{\partial t} = \frac{1}{\sqrt{N}} [f_-(x) - f_+(x)] \frac{\partial \Pi}{\partial \xi} \\ + \frac{1}{2} \frac{1}{N} [f_-(x) + f_+(x)] \frac{\partial^2 \Pi}{\partial \xi^2} \\ + \frac{1}{N} [f'_-(x) - f'_+(x)] \frac{\partial}{\partial \xi} [\xi P] + \dots, \quad (92)$$

where the functions  $f_-(x)$  and  $f_+(x)$  are given by:

$$f_-(x) = cx^2 + dx, \quad f_+(x) = 2bx(1-x). \quad (93)$$



For the left- and right-hand sides to balance in Eq. (92), a rescaled time  $\tau = t/N$  needs to be introduced. Then to leading order one finds  $dx/d\tau = f(x)$ , where  $f(x) = f_+(x) - f_-(x)$ . At next to leading order Eq. (91) is found with  $g(x) = f_+(x) + f_-(x)$ . The explicit form of the macroscopic equation is

$$\frac{dx}{d\tau} = x(r - ax), \quad (94)$$

where  $r = 2b - d$  and  $a = 2b + c$ . Equation (94) is the logistic equation, which is the usual phenomenological way to model intraspecies competition.

Since the Fokker–Planck equation (91) describes a linear process, its solution is a Gaussian. This means that the probability distribution  $\Pi(\xi, t)$  is completely specified by the first two moments  $\langle \xi(t) \rangle$  and  $\langle \xi^2(t) \rangle$ . Multiplying Eq. (91) by  $\xi$  and  $\xi^2$  and integrating over all  $\xi$  one finds

$$\begin{aligned} \frac{d}{dt} \langle \xi(t) \rangle &= f'(x_M(t)) \langle \xi(t) \rangle, \\ \frac{d}{dt} \langle \xi^2(t) \rangle &= 2f'(x_M(t)) \langle \xi^2(t) \rangle + g(x_M(t)). \end{aligned} \quad (95)$$

We have chosen the initial condition to be  $x_0 = n_0/N$ , which implies that  $\xi(0) = 0$ . The first equation in (95) then implies that  $\langle \xi(t) \rangle = 0$  for all  $t$ . Multiplying the second equation by  $f^{-2}(x_M(t))$  one finds that

$$\langle \xi^2(t) \rangle = f^2(x_M(t)) \int_0^t dt' \frac{g(x_M(t'))}{f^2(x_M(t'))}, \quad (96)$$

and so the determination of  $\langle \xi^2(t) \rangle$  is reduced to quadrature. The method can be applied to systems with more than one stochastic variable and those which are not one-step processes. Details are given in van Kampen's book [53].

## Future Directions

This article has focused largely on classical topics in the theory of stochastic processes, since these form the foundations on which the subject is built. Much of the current work, and one would expect future work, will be numerical in character. Some of this will begin from a basic Markovian description in the form of reactions among chemical species – even if the system is not chemical in nature (Example 3 of Sect. “The Master Equation” is an example). A straightforward algorithm developed by Gillespie [17,18], and since then extended and improved [3,19], provides an efficient way of simulating such systems. It thus provides a valuable method of investigating systems

which may be formulated as complicated multivariable master equations, which complements the methods we have discussed here. However, many current studies do not begin from a system which can be described in this way, and there is every indication that this will be more true in the future. For instance, in agent based models the stochastic element may be due to mutations in characteristics, traits or behavior, which may be difficult or impossible to formulate mathematically. Such agent based models are certainly individually based, but each individual may have different attributes and generally behave in such a complex way that only numerical simulations can be used to explore the behavior of the system as a whole. Although these complex systems may be used to model more realistic situations, the well-known problems associated with the large number of parameters typically required to describe such systems, will mean that simplified versions will need to be analyzed in order to understand them at a deeper level. These simpler models are likely to include those where the agents of a particular species are essentially identical. In this article we have discussed how the classical equations of the theory of stochastic processes, such as the Fokker–Planck equation, can be obtained from such models. They will therefore form a bridge between the agent-based approaches which are expected to become more prevalent in the future, and the analytic approaches which lie at the heart of the theory of stochastic processes.

## Bibliography

### Primary Literature

1. Abramowitz M, Stegun I (Eds) (1965) Handbook of mathematical functions. Dover, New York
2. Bachelier L (1900) Théorie de la spéculation. Annales Scientifiques de L'Ecole Normale Supérieure III(17):21–86
3. Cao Y, Li H, Petzold L (2004) Efficient formulation of the stochastic simulation algorithm for chemically reacting systems. J Chem Phys 121:4059–4067
4. Chandrasekhar S (1943) Stochastic problems in physics and astronomy. Rev Mod Phys 15:1–89. Reprinted in Wax (1954)
5. Cox DR, Miller HD (1968) The theory of stochastic processes, Chap 3. Chapman and Hall, London
6. Crow JF, Kimura M (1970) An introduction to population genetics theory. Harper and Row, New York
7. Ehrenfest P, Ehrenfest T (1907) Über zwei bekannte Einwände gegen das Boltzmannsche H-Theorem. Phys Z 8:311–314
8. Einstein A (1905) Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. Ann Physik 17:549–560. For a translation see: A. Einstein, “Investigations on the Theory of the Brownian Movement” Fürth R (ed), Cowper AD (tr) (Dover, New York, 1956). Chapter I

9. Einstein A (1906) Zur Theorie der Brownschen Bewegung. *Ann Physik* 19:371–381. For a translation see: A. Einstein, “Investigations on the Theory of the Brownian Movement” Fürth R (ed), Cowper AD (tr) (Dover, New York, 1956). Chapter II
10. Feller W (1968) An introduction to probability theory and its applications, Chap XV, 3rd edn. Wiley, New York
11. Feynman RP (1948) Space-time approach to non-relativistic quantum mechanics. *Rev Mod Phys* 20:367–387
12. Feynman RP, Hibbs AR (1965) Quantum mechanics and path integrals, Chap 12. McGraw-Hill, New York
13. Fisher RA (1930) The genetical theory of natural selection. Clarendon Press, Oxford
14. Fokker AD (1914) Die mittlere Energie rotierende elektrischer Dipole im Strahlungsfeld. *Ann Physik* 43:810–820
15. Gantmacher FR (1959) The theory of matrices, Chap 13, Sect 6, vol 2. Chelsea Publishing Co., New York
16. Gardiner CW (2004) Handbook of stochastic methods, 3rd edn. Springer, Berlin
17. Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J Comput Phys* 22:403–434
18. Gillespie DT (1977) Exact stochastic simulation of coupled chemical reactions. *J Phys Chem* 81:2340–2361
19. Gillespie DT (2001) Approximate accelerated stochastic simulation of chemically reacting systems. *J Chem Phys* 115:1716–1733
20. Graham R (1977) Path integral formulation of general diffusion processes. *Z Physik B* 26:281–290
21. Graham R (1975) Macroscopic theory of fluctuations and instabilities. In: Riste T (ed) *Fluctuations, Instabilities, and Phase Transitions*. Plenum, New York, pp 215–293
22. Haken H (1983) *Synergetics*. Springer, Berlin. Sect 6.3
23. Kac M (1947) Random walk and the theory of Brownian motion. *Amer Math Mon* 54:369–391
24. Karlin S, McGregor J (1964) On some stochastic models in genetics. In: Gurland J (ed) *Stochastic problems in medicine and biology*. University of Wisconsin Press, Madison, pp 245–279
25. Kendall DG (1948) On some modes of population growth leading to R. A. Fisher’s logarithmic series distribution. *Biometrika* 35:6–15
26. Kolmogorov AN (1931) Über die analytischen Methoden in der Wahrscheinlichkeitsrechnung. *Math Ann* 104:415–458
27. Kolmogorov AN (1936) Anfangsgründe der Theorie der Markoffschen Ketten mit unendlich vielen möglichen Zuständen. *Mat Sbornik (N.S.)* 1:607–610
28. Krafft O, Schaefer M (1993) Mean passage times for tridiagonal transition matrices and a two-parameter Ehrenfest urn model. *J Appl Prob* 30:964–970
29. Kramers HA (1940) Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica* 7:284–304
30. Langevin P (1908) Sur la théorie du mouvement brownien. *C R Acad Sci Paris* 146:530–533. For a translation see: D. S. Lemons and A. Gythiel, *Am. J. Phys.* 65: 1079–1081 (1997)
31. Malécot G (1944) Sur un problème de probabilités en chaîne que pose la génétique. *C R Acad Sci Paris* 219:379–381
32. Markov AA (1906) Rasprostranenie zakona bol’shih chisel na velichiny, zavisyaschie drug ot druga. *Izv Fiz-Matem Obsch Kazan Univ (Series 2)* 15:135–156. See also: “Extension of the limit theorems of probability theory to a sum of variables connected in a chain”, in Appendix B of R. Howard “Dynamic Probabilistic Systems, vol 1: Markov Chains” (John Wiley and Sons, 1971)
33. May RM (1973) *Model ecosystems*, Chap 5. Princeton University Press, Princeton
34. McKane AJ, Newman TJ (2004) Stochastic models in population biology and their deterministic analogs. *Phys Rev E* 70:041902
35. McKane AJ, Newman TJ (2005) Predator-prey cycles from resonant amplification of demographic stochasticity. *Phys Rev Lett* 94:218102
36. Moran PAP (1958) Random processes in genetics. *Proc Cambridge Philos Soc* 54:60–72
37. Moran PAP (1962) The statistical processes of evolutionary theory, Chap 4. Clarendon Press, Oxford
38. Moyal JE (1949) Stochastic processes and statistical physics. *J Roy Stat Soc (London)* B 11:150–210
39. Nordsieck A, Lamb WE Jr, Uhlenbeck GE (1940) On the theory of cosmic-ray showers. I. The Furry model and the fluctuation problem. *Physica* 7:344–360
40. Onsager L, Machlup S (1953) Fluctuations and irreversible processes. *Phys Rev* 91:1505–1512
41. Pauli W (1928) Probleme der modernen Physik. In: Debye P (ed) *Festschrift zum 60. Geburtstag A. Sommerfeld*. Hirzel, Leipzig, p 30
42. Pearson K (1905) The problem of the random walk. *Nature* 72:294,342
43. Planck M (1917) Über einen Satz der statistischen Dynamik und seine Erweiterung in der Quantentheorie. *Abh Preuss Akad Wiss Berl* 24:324–341
44. Rayleigh L (1891) Dynamical problems in illustration of the theory of gases. *Phil Mag* 32:424–445
45. Reichl LE (1998) A modern course in statistical physics, Chap 5, 2nd edn. Wiley, New York
46. Risken H (1989) The Fokker-Planck equation, 2nd edn. Springer, Berlin
47. Schiff LI (1968) Quantum mechanics, Chap 4, 3rd edn. McGraw-Hill, Tokyo
48. Schulman LS (1981) Techniques and applications of path-integration, Chap 5. Wiley, New York
49. Siegert AJF (1949) On the approach to statistical equilibrium. *Phys Rev* 76:1708–1714
50. Sneddon IN (1957) Elements of partial differential equations, Chap 2. McGraw-Hill, New York
51. van Kampen NG (1961) A power series expansion of the master equation. *Can J Phys* 39:551–567
52. van Kampen NG (1981) Itô versus Stratonovich. *J Stat Phys* 24:175–187
53. van Kampen NG (1992) Stochastic processes in physics and chemistry, 2nd edn. North-Holland, Amsterdam
54. von Smoluchowski M (1906) Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen. *Ann Physik* 21:756–780
55. von Smoluchowski M (1916) Drei Vorträge über Diffusion, Brownsche Bewegung, und Koagulation von Kolloidteilchen. *Phys Z* 17:571–599
56. Wiener N (1921) The average of an analytic functional. *Proc Natl Acad Sci USA* 7:253–260
57. Wiener N (1921) The average of an analytic functional and the Brownian movement. *Proc Natl Acad Sci USA* 7:294–298

58. Wright S (1931) Evolution in Mendelian populations. *Genetics* 16:97–159
59. Zinn-Justin J (2002) Quantum field theory and critical phenomena, 4th edn. Clarendon Press, Oxford. Sect 4.8.2
60. Zwanzig R (1973) Nonlinear generalized Langevin equations. *J Stat Phys* 9:215–220

## Books and Reviews

- Wax N (1954) Selected papers on noise and stochastic processes. Dover, New York

## Stochastic Volatility

TORBEN G. ANDERSEN<sup>1,2,3</sup>, LUCA BENZONI<sup>4</sup>

<sup>1</sup> Kellogg School of Management, Northwestern University, Evanston, USA

<sup>2</sup> NBER, Cambridge, USA

<sup>3</sup> CREATES, Aarhus, Denmark

<sup>4</sup> Federal Reserve Bank of Chicago, Chicago, USA

## Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Model Specification](#)

[Realized Volatility](#)

[Applications](#)

[Estimation Methods](#)

[Future Directions](#)

[Acknowledgments](#)

[Bibliography](#)

## Glossary

**Implied volatility** The value of asset return volatility which equates a model-implied derivative price to the observed market price. Most notably, the term is used to identify the volatility implied by the Black and Scholes [63] option pricing formula.

**Quadratic return variation** The ex-post sample-path return variation over a fixed time interval.

**Realized volatility** The sum of finely sampled squared asset return realizations over a fixed time interval. It is an estimate of the quadratic return variation over such time interval.

**Stochastic volatility** A process in which the return variation dynamics include an unobservable shock which cannot be predicted using current available information.

## Definition of the Subject

Given the importance of return volatility on a number of practical financial management decisions, the efforts to provide good real-time estimates and forecasts of current and future volatility have been extensive. The main framework used in this context involves stochastic volatility models. In a broad sense, this model class includes GARCH, but we focus on a narrower set of specifications in which volatility follows its own random process, as is common in models originating within financial economics. The distinguishing feature of these specifications is that volatility, being inherently unobservable and subject to independent random shocks, is not measurable with respect to observable information. In what follows, we refer to these models as *genuine* stochastic volatility models.

Much modern asset pricing theory is built on continuous-time models. The natural concept of volatility within this setting is that of genuine stochastic volatility. For example, stochastic volatility (jump-)diffusions have provided a useful tool for a wide range of applications, including the pricing of options and other derivatives, the modeling of the term structure of risk-free interest rates, and the pricing of foreign currencies and defaultable bonds. The increased use of intraday transaction data for construction of so-called realized volatility measures provides additional impetus for considering genuine stochastic volatility models. As we demonstrate below, the realized volatility approach is closely associated with the continuous-time stochastic volatility framework of financial economics.

There are some unique challenges in dealing with genuine stochastic volatility models. For example, volatility is truly latent and this feature complicates estimation and inference. Further, the presence of an additional state variable – volatility – renders the model less tractable from an analytic perspective. We review how such challenges have been addressed through development of new estimation methods and imposition of model restrictions allowing for closed-form solutions while remaining consistent with the dominant empirical features of the data.

## Introduction

The label Stochastic Volatility is applied in two distinct ways in the literature. For one, it is used to signify that the (absolute) size of the innovations of a time series displays random fluctuations over time. Descriptive models of financial time series almost invariably embed this feature nowadays as asset return series tend to display alternating quiet and turbulent periods of varying length and intensity. To distinguish this feature from models that operate with an a priori known or deterministic path

for the volatility process, the random evolution of the conditional return variance is termed stochastic volatility. The simplest case of deterministic volatility is the constant variance assumption invoked in, e.g., the Black and Scholes [63] framework. Another example is modeling the variance purely as a given function of calendar time, allowing only for effects such as time-of-year (seasonals), day-of-week (institutional and announcement driven) or time-of-day (diurnal effects due to, e.g., market microstructure features). Any model not falling within this class is then a stochastic volatility model. For example, in the one-factor continuous-time Cox, Ingersoll, and Ross [113] (CIR) model the (stochastic) level of the short term interest rate governs the dynamics of the (instantaneous) drift and diffusion term of all zero-coupon yields. Likewise, in GARCH models the past return innovations govern the one-period ahead conditional mean and variance. In both models, the volatility is known, or deterministic, at a given point in time, but the random evolution of the processes renders volatility stochastic for any horizon beyond the present period.

The second notion of stochastic volatility, which we adopt henceforth, refers to models in which the return variation dynamics is subject to an unobserved random shock so that the volatility is inherently latent. That is, the current volatility state is not known for sure, conditional on the true data generating process and the past history of all available discretely sampled data. Since the CIR and GARCH models described above render the current (conditional) volatility known, they are not stochastic volatility models in this sense. In order to make the distinction clear cut, we follow Andersen [10] and label this second, more restrictive, set *genuine* stochastic volatility (SV) models.

There are two main advantages to focusing on SV models. First, much asset pricing theory is built on continuous-time models. Within this class, SV models tend to fit more naturally with a wide array of applications, including the pricing of currencies, options, and other derivatives, as well as the modeling of the term structure of interest rates. Second, the increasing use of high-frequency intraday data for construction of so-called realized volatility measures is also starting to push the GARCH models out of the limelight as the realized volatility approach is naturally linked to the continuous-time SV framework of financial economics.

One drawback is that volatility is not measurable with respect to observable (past) information in the SV setting. As such, an estimate of the current volatility state must be filtered out from a noisy environment and the estimate will change as future observations become available. Hence, in-sample estimation typically involves smoothing

techniques, not just filtering. In contrast, the conditional variance in GARCH is observable given past information, which renders (quasi-)maximum likelihood techniques for inference quite straightforward while smoothing techniques have no role. As such, GARCH models are easier to estimate and practitioners often rely on them for time-series forecasts of volatility. However, the development of powerful method of simulated moments, Markov Chain Monte Carlo (MCMC) and other simulation based procedures for estimation and forecasting of SV models may well render them competitive with ARCH over time on that dimension.

Direct indications of the relations between SV and GARCH models are evident in the sequence of papers by Dan Nelson and Dean Foster exploring the SV diffusion limits of ARCH models as the case of continuous sampling is approached, see, e.g., Nelson and Foster [219]. Moreover, as explained in further detail in the estimation section below, it can be useful to summarize the dynamic features of asset returns by tractable pseudo-likelihood scores obtained from GARCH-style models when performing simulation based inference for SV models. As such, the SV and GARCH frameworks are closely related and should be viewed as complements. Despite these connections we focus, for the sake of brevity, almost exclusively on SV models and refer the interested reader to the GARCH chapter for further information.

The literature on SV models is vast and rapidly growing, and excellent surveys are available, e.g., Ghysels et al. [158] and Shephard [239,240]. Consequently, we focus on providing an overview of the main approaches with illustrations of the scope for applications of these models to practical finance problems.

## Model Specification

The original econometric studies of SV models were invariably cast in discrete time and they were quite similar in structure to ARCH models, although endowed with a more explicit structural interpretation. Recent work in the area has been mostly directly towards a continuous time setting and motivated by the typical specifications in financial economics. This section briefly reviews the two alternative approaches to specification of SV models.

### Discrete-Time SV Models and the Mixture-of-Distributions Hypothesis

Asset pricing theory contends that financial asset prices reflect the discounted value of future expected cash flows, implying that all news relevant for either discount rates or cash flows should induce a shift in market prices. Since

economic news items appear almost continuously in real time, this perspective rationalizes the ever-changing nature of prices observed in financial markets. The process linking news arrivals to price changes may be complex, but if it is stationary in the statistical sense it will nonetheless produce a robust theoretical association between news arrivals, market activity and return volatility. In fact, if the number of news arrival is very large, standard central limit theory will tend to imply that asset returns are approximately normally distributed *conditional* on the news count. More generally, variables such as the trading volume, the number of transactions or the number of price quotes are also naturally related to the intensity of the information flow. This line of reasoning has motivated specifications such as

$$y_t | s_t \rightsquigarrow N(\mu_y s_t, \sigma_y^2 s_t), \quad (1)$$

where  $y_t$  is an “activity” variable related to the information flow,  $s_t$  is a positive intensity process reflecting the rate of news arrivals,  $\mu_y$  represents the mean response to an information event, and  $\sigma_y$  is a pure scaling parameter.

This is a normal mixture model, where the  $s_t$  process governs or “mixes” the scale of the distribution across the periods. If  $s_t$  is constant, this is simply an i.i.d. Gaussian process for returns and possible other related variables. However, this is clearly at odds with the empirical evidence for, e.g., return volatility and trading volume. Therefore,  $s_t$  is typically stipulated to follow a separate stochastic process with random innovations. Hence, each period the return series is subject to two separate shocks, namely the usual idiosyncratic error term associated with the (normal) return distribution, but also a shock to the variance or volatility process,  $s_t$ . This endows the return process with genuine stochastic volatility, reflecting the random intensity of news arrivals. Moreover, it is typically assumed that only returns, transactions and quotes are observable, but not the actual value of  $s_t$  itself, implying that  $\sigma_y$  cannot be separately identified. Hence, we simply fix this parameter at unity.

The time variation in the information flow series induces a fat-tailed unconditional distribution, consistent with stylized facts for financial return and, e.g., trading volume series. Intuitively, days with a lot of news display more rapid price fluctuations and trading activity than days with a low news count. In addition, if the  $s_t$  process is positively correlated, then shocks to the conditional mean and variance processes for  $y_t$  will be persistent. This is consistent with the observed clustering in financial markets, where return volatility and trading activity are contemporaneously correlated and each display pronounced positive serial dependence.

The inherent randomness and unobserved nature of the news arrival process, even during period  $t$ , renders the true mean and variance series latent. This property is the major difference with the GARCH model class, in which the one-step-ahead conditional mean and variance are a known function of observed variables at time  $t - 1$ . As such, for genuine SV models, we must distinguish the full, but infeasible, information set ( $s_t \in \mathcal{F}_t$ ) and the observable information set ( $s_t \notin \mathcal{I}_t$ ). This basic latency of the mixing variable (state vector) of the SV model complicates inference and forecasting procedures as discussed below.

For short horizon returns,  $\mu_y$  is nearly negligible and can reasonably be ignored or simply fixed at a small constant value, and the series can then be demeaned. This simplification produces the following return (innovation) model,

$$r_t = \sqrt{s_t} z_t, \quad (2)$$

where  $z_t$  is an i.i.d. standard normal variable, implying a simple normal-mixture representation,

$$r_t | s_t \rightsquigarrow N(0, s_t). \quad (3)$$

Univariate return models of the form (3) as well as multivariate systems including a return variable along with other related market activity variables, such as the transactions count, the quote intensity or the aggregate trading volume, stem from the Mixture-of-Distributions Hypothesis (MDH).

Actual implementation of the MDH hinges on a particular representation of the information-arrival process  $s_t$ . Clark [102] uses trading volume as a proxy for the activity variable, a choice motivated by the high contemporaneous correlation between return volatility and volume. Tauchen and Pitts [247] follow a structural approach to characterize the joint distribution of the daily return and volume relation governed by the underlying latent information flow  $s_t$ . However, both these models assume temporal independence of the information flow, thus failing to capture the clustering in these series. Partly in response, Gallant et al. [153] examine the joint conditional return-volume distribution without imposing any structural MDH restrictions. Nonetheless, many of the original discrete-time SV specifications are compatible with the MDH framework, including Taylor [249]<sup>1</sup>, who proposes an autoregressive parametrization of the latent log-volatility (or information flow) variable

$$\log(s_{t+1}) = \eta_0 + \eta_1 \log(s_t) + u_t, \quad u_t \rightsquigarrow \text{i.i.d.}(0, \sigma_u^2), \quad (4)$$

<sup>1</sup>Discrete-time SV models go father back in time, at least to the early paper by Rosenberg [232] recently reprinted in Shephard [240].



where the error term,  $u_t$ , may be correlated with the disturbance term,  $z_t$ , in the return Eq. (2) so that  $\rho = \text{corr}(u_t, z_t) \neq 0$ . If  $\rho < 0$ , downward movements in asset prices result in higher future volatility as also predicted by the so-called ‘leverage effect’ in the exponential GARCH, or EGARCH, form of Nelson [218] and the asymmetric GARCH model of Glosten et al. [160].

Early tests of the MDH include Lamoureux and Lastrapes [194] and Richardson and Smith [231]. Subsequently, Andersen [11] studies a modified version of the MDH that provides a much improved fit to the data. Further refinements of the MDH specification have been pursued by, e.g., Liesenfeld [198,199] and Bollerslev and Jubinsky [67]. Among the first empirical studies of the related approach of stochastic time changes are Ané and Geman [29], who focus on stock returns, and Conley et al. [109], who focus on the short-term risk-free interest rate.

### Continuous-Time Stochastic Volatility Models

Asset returns typically contain a predictable component, which compensates the investor for the risk of holding the security, and an unobservable shock term, which cannot be predicted using current available information. The conditional asset return variance pertains to the variability of the unobservable shock term. As such, over a non-infinitesimal horizon it is necessary to first specify the conditional mean return (e.g., through an asset pricing model) in order to identify the conditional return variation. In contrast, over an infinitesimal time interval this is not necessary because the requirement that market prices do not admit arbitrage opportunities implies that return innovations are an order of magnitude larger than the mean return. This result has important implications for the approach we use to model and measure volatility in continuous time.

Consider an asset with log-price process  $\{p(t), t \in [0, T]\}$  defined on a probability space  $(\Omega, \mathcal{F}, P)$ . Following Andersen et al. [19] we define the continuously compounded asset return over a time interval from  $t-h$  to  $t$ ,  $0 \leq h \leq t \leq T$ , to be

$$r(t, h) = p(t) - p(t-h). \quad (5)$$

A special case of (5) is the cumulative return up to time  $t$ , which we denote  $r(t) \equiv r(t, t) = p(t) - p(0)$ ,  $0 \leq t \leq T$ . Assume the asset trades in a frictionless market void of arbitrage opportunities and the number of potential discontinuities (jumps) in the price process per unit time is finite. Then the log-price process  $p$  is a semi-martingale (e.g., Back [33]) and therefore the cumulative return  $r(t)$

admits the decomposition (e.g., Protter [229])

$$r(t) = \mu(t) + M^C(t) + M^J(t), \quad (6)$$

where  $\mu(t)$  is a predictable and finite variation process,  $M^C(t)$  a continuous-path infinite-variation martingale, and  $M^J(t)$  is a compensated finite activity jump martingale. Over a discrete time interval the decomposition (6) becomes

$$r(t, h) = \mu(t, h) + M^C(t, h) + M^J(t, h), \quad (7)$$

where  $\mu(t, h) = \mu(t) - \mu(t-h)$ ,  $M^C(t, h) = M^C(t) - M^C(t-h)$ , and  $M^J(t, h) = M^J(t) - M^J(t-h)$ .

Denote now with  $[r, r]$  the quadratic variation of the semi-martingale process  $r$ , where (Protter [229])

$$[r, r]_t = r(t)^2 - 2 \int r(s-) dr(s), \quad (8)$$

and  $r(t-) = \lim_{s \uparrow t} r(s)$ . If the finite variation process  $\mu$  is continuous, then its quadratic variation is identically zero and the predictable component  $\mu$  in decomposition (7) does not affect the quadratic variation of the return  $r$ . Thus, we obtain an expression for the quadratic return variation over the time interval from  $t-h$  to  $t$ ,  $0 \leq h \leq t \leq T$  (e.g., Andersen et al. [21] and Barndorff-Nielsen and Shephard [51,52]):

$$\begin{aligned} \text{QV}(t, h) &= [r, r]_t - [r, r]_{t-h} \\ &= [M^C, M^C]_t - [M^C, M^C]_{t-h} \\ &\quad + \sum_{t-h < s \leq t} \Delta M^2(s) \\ &= [M^C, M^C]_t - [M^C, M^C]_{t-h} \\ &\quad + \sum_{t-h < s \leq t} \Delta r^2(s). \end{aligned} \quad (9)$$

Most continuous-time models for asset returns can be cast within the general setting of Eq. (7), and Eq. (9) provides a framework to study the model-implied return variance. For instance, the Black and Scholes [63] model is a special case of the setting described by Eq. (7) in which the conditional mean process  $\mu$  is constant, the continuous martingale  $M^C$  is a standard Brownian motion process, and the jump martingale  $M^J$  is identically zero:

$$dp(t) = \mu dt + \sigma dW(t). \quad (10)$$

In this case, the quadratic return variation over the time interval from  $t-h$  to  $t$ ,  $0 \leq h \leq t \leq T$ , simplifies to

$$\text{QV}(t, h) = \int_{t-h}^t \sigma^2 ds = \sigma^2 h, \quad (11)$$

that is, return volatility is constant over any time interval of length  $h$ .

A second notable example is the jump-diffusion model of Merton [214],

$$dp(t) = (\mu - \lambda \bar{\xi})dt + \sigma dW(t) + \xi(t)dq_t, \quad (12)$$

where  $q$  is a Poisson process uncorrelated with  $W$  and governed by the constant jump intensity  $\lambda$ , i. e.,  $\text{Prob}(dq_t = 1) = \lambda dt$ . The scaling factor  $\xi(t)$  denotes the magnitude of the jump in the return process if a jump occurs at time  $t$ . It is assumed to be normally distributed,

$$\xi(t) \rightsquigarrow N(\bar{\xi}, \sigma_{\xi}^2). \quad (13)$$

In this case, the quadratic return variation process over the time interval from  $t-h$  to  $t$ ,  $0 \leq h \leq t \leq T$  becomes

$$\begin{aligned} QV(t, h) &= \int_{t-h}^t \sigma^2 ds + \sum_{t-h \leq s \leq t} J(s)^2 \\ &= \sigma^2 h + \sum_{t-h \leq s \leq t} J(s)^2, \end{aligned} \quad (14)$$

where  $J(t) \equiv \xi(t)dq(t)$  is non-zero only if a jump actually occurs.

Finally, a broad class of stochastic volatility models is defined by

$$dp(t) = \mu(t)dt + \sigma(t)dW(t) + \xi(t)dq_t, \quad (15)$$

where  $q$  is a constant-intensity Poisson process with log-normal jump amplitude (13). Equation (15) is also a special case of (7) and the associated quadratic return variation over the time interval from  $t-h$  to  $t$ ,  $0 \leq h \leq t \leq T$ , is

$$\begin{aligned} QV(t, h) &= \int_{t-h}^t \sigma(s)^2 ds + \sum_{t-h \leq s \leq t} J(s)^2 \\ &\equiv IV(t, h) + \sum_{t-h \leq s \leq t} J(s)^2. \end{aligned} \quad (16)$$

As in the general case of Eq. (9), Eq. (16) identifies the contribution of diffusive volatility, termed ‘integrated variance’ (IV), and cumulative squared jumps to the total quadratic variation.

Early applications typically ignored jumps and focused exclusively on the integrated variance component. For instance, IV plays a key role in Hull and White’s [174] SV option pricing model, which we discuss in Sect. “Options” below along with other option pricing applications. For illustration, we focus here on the SV model specification by Wiggins [256]:

$$dp(t) = \mu dt + \sigma(t)dW_p(t) \quad (17)$$

$$d\sigma(t) = f(\sigma(t))dt + \eta\sigma(t)dW_{\sigma}(t), \quad (18)$$

where the innovations to the return  $dp$  and volatility  $\sigma$ ,  $W_p$  and  $W_{\sigma}$ , are standard Brownian motions. If we define  $y = \log(\sigma)$  and apply Itô’s formula we obtain

$$\begin{aligned} dy(t) &= d\log(\sigma(t)) \\ &= \left[ -\frac{1}{2}\eta^2 + \frac{f(\sigma(t))}{\sigma(t)} \right] dt + \eta dW_{\sigma}(t). \end{aligned} \quad (19)$$

Wiggins approximates the drift term  $f(\sigma(t)) \approx \{\alpha + \kappa[\log(\bar{\sigma}) - \log(\sigma(t))]\}\sigma(t)$ . Substitution in Eq. (19) yields

$$d\log(\sigma(t)) = [\bar{\alpha} - \kappa \log(\sigma(t))]dt + \eta dW_{\sigma}(t), \quad (20)$$

where  $\bar{\alpha} = \alpha + \kappa \log(\bar{\sigma}) - \frac{1}{2}\eta^2$ . As such, the logarithmic standard deviation process in Wiggins has diffusion dynamics similar in spirit to Taylor’s discrete time AR(1) model for the logarithmic information process, Eq. (4). As in Taylor’s model, negative correlation between return and volatility innovations,  $\rho = \text{corr}(W_p, W_{\sigma}) < 0$ , generates an asymmetric response of volatility to return shocks similar to the leverage effect in discrete-time EGARCH models.

More recently, several authors have imposed restrictions on the continuous-time SV jump-diffusion (15) that render the model more tractable while remaining consistent with the empirical features of the data. We return to these models in Sect. “Options” below.

## Realized Volatility

Model-free measures of return variation constructed only from concurrent return realizations have been considered at least since Merton [215]. French et al. [148] construct monthly historical volatility estimates from daily return observations. More recently, the increased availability of transaction data has made it possible to refine early measures of historical volatility into the notion of ‘realized volatility’, which is endowed with a formal theoretical justification as an estimator of the quadratic return variation as first noted in Andersen and Bollerslev [18]. The realized volatility of an asset return  $r$  over the time interval from  $t-h$  to  $t$  is

$$RV(t, h; n) = \sum_{i=1}^n r \left( t-h + \frac{ih}{n}, \frac{h}{n} \right)^2. \quad (21)$$

Semi-martingale theory ensures that the realized volatility measure  $RV$  converges to the return quadratic variation  $QV$ , previously defined in Eq. (9), when the sampling frequency  $n$  increases. We point the interested reader to, e. g., Andersen et al. [19] to find formal arguments in support of

this claim. Here we convey intuition for this result by considering the special case in which the asset return follows a continuous-time diffusion without jumps,

$$dp(t) = \mu(t)dt + \sigma(t)dW(t). \quad (22)$$

As in Eq. (21), consider a partition of the  $[t-h, t]$  interval with mesh  $h/n$ . A discretization of the diffusion (22) over a sub-interval from  $(t-h + (i-1)h/n)$  to  $(t-h + ih/n)$ ,  $i = 1, \dots, n$ , yields

$$r\left(t-h + \frac{ih}{n}, \frac{h}{n}\right) \approx \mu\left(t-h + \frac{(i-1)h}{n}\right) \frac{h}{n} + \sigma\left(t-h + \frac{(i-1)h}{n}\right) \Delta W\left(t-h + \frac{ih}{n}\right), \quad (23)$$

where  $\Delta W(t-h + ih/n) = W(t-h + ih/n) - W(t-h + (i-1)h/n)$ .

Suppressing time indices, the squared return  $r^2$  over the time interval of length  $h/n$  is therefore:

$$r^2 = \mu^2 \left(\frac{h}{n}\right)^2 + 2\mu\sigma\Delta W\left(\frac{h}{n}\right) + \sigma^2(\Delta W)^2. \quad (24)$$

As  $n \rightarrow \infty$  the first two terms vanish at a rate higher than the last one. In particular, to a first order approximation the squared return equals the squared return innovation and therefore the squared return conditional mean and variance are

$$E[r^2|\mathcal{F}_t] \approx \sigma^2 \frac{h}{n} \quad (25)$$

$$\text{Var}[r^2|\mathcal{F}_t] \approx 2\sigma^4 \left(\frac{h}{n}\right)^2. \quad (26)$$

The no-arbitrage condition implies that return innovations are serially uncorrelated. Thus, summing over  $i = 1, \dots, n$  we obtain

$$\begin{aligned} E[RV(t, h, n)|\mathcal{F}_t] &= \sum_{i=1}^n E\left[r\left(t-h + \frac{ih}{n}, \frac{h}{n}\right)^2 | \mathcal{F}_t\right] \\ &\approx \sum_{i=1}^n \sigma^2\left(t-h + \frac{(i-1)h}{n}\right)^2 \frac{h}{n} \\ &\approx \int_{t-h}^t \sigma(s)^2 ds \end{aligned} \quad (27)$$

$$\begin{aligned} \text{Var}[RV(t, h, n)|\mathcal{F}_t] &= \sum_{i=1}^n \text{Var}\left[r\left(t-h + \frac{ih}{n}, \frac{h}{n}\right)^2 | \mathcal{F}_t\right] \\ &\approx \sum_{i=1}^n 2\sigma^4\left(t-h + \frac{(i-1)h}{n}\right)^4 \left(\frac{h}{n}\right)^2 \\ &\approx 2\left(\frac{h}{n}\right) \int_{t-h}^t \sigma(s)^4 ds. \end{aligned} \quad (28)$$

Equation (27) illustrates that realized volatility is an unbiased estimator of the return quadratic variation, while Eq. (28) shows that the estimator is consistent as its variance shrinks to zero when we increase the sampling frequency  $n$  and keep the time interval  $h$  fixed. Taken together, these results suggest that RV is a powerful and model-free measure of the return quadratic variation. Effectively, RV gives practical empirical content to the latent volatility state variable underlying the models previously discussed in Sect. “Continuous-Time Stochastic Volatility Models”.

Two issues complicate the practical application of the convergence results illustrated in Eqs. (27) and (28). First, a continuum of instantaneous return observations must be used for the conditional variance in Eq. (28) to vanish. In practice, only a discrete price record is observed, and thus an inevitable discretization error is present. Barndorff-Nielsen and Shephard [52] develop an asymptotic theory to assess the effect of this error on the RV estimate (see also [209]). Second, market microstructure effects (e.g., price discreteness, bid-ask spread positioning due to dealer inventory control, and bid-ask bounce) contaminate the return observations, especially at the ultra-high frequency. These effects tend to generate spurious correlations in the return series which can be partially eliminated by filtering the data prior to forming the RV estimates. However, this strategy is not a panacea and much current work studies the optimal sampling scheme and the construction of improved realized volatility in the presence of microstructure noise. This growing literature is surveyed by Hansen and Lunde [165], Bandi and Russell [46], McAleer and Medeiros [205], and Andersen and Benzoni [14]. Recent notable contributions to this literature include Bandi and Russell [45], Barndorff-Nielsen et al. [49], Diebold and Strasser [121], and Zhang, Mykland, and Aït-Sahalia [262]. Related, there is the issue of how to construct RV measures when the market is rather illiquid. One approach is to use a lower sampling frequency and focus on longer-horizon RV measure. Alternatively the literature has explored volatility measures that are more robust to situations in which the noise-to-

signal ratio is high, e. g., Alizadeh et al. [8], Brandt and Diebold [72], Brandt and Jones [73], Gallant et al. [151], Garman and Klass [157], Parkinson [221], Schwert [237], and Yang and Zhang [259] consider the high-low price range measure. Dobrev [122] generalizes the range estimator to high-frequency data and shows its link with RV measures.

Equations (27) and (28) also underscore an important difference between RV and other volatility measures. RV is an ex-post model-free estimate of the quadratic variation process. This is in contrast to ex-ante measures which attempt to forecast future quadratic variation using information up to current time. The latter class includes parametric GARCH-type volatility forecasts as well as forecasts built from stochastic volatility models through, e. g., the Kalman filter (e. g., [167,168]), the particle filter (e. g., [186,187]) or the reprojection method (e. g., [152,155]).

More recently, other studies have pursued more direct time-series modeling of volatility to obtain alternative ex-ante forecasts. For instance, Andersen et al. [21] follow an ARMA-style approach, extended to allow for long memory features, to model the logarithmic foreign exchange rate realized volatility. They find the fit to dominate that of traditional GARCH-type models estimated from daily data. In a related development, Andersen, Bollerslev, and Meddahi [24,25] exploit the general class of Eigenfunction Stochastic Volatility (ESV) models introduced by Meddahi [208] to provide optimal analytic forecast formulas for realized volatility as a function of past realized volatility. Other scholars have pursued more general model specifications to improve forecasting performance. Ghysels et al. [159] consider Mixed Data Sampling (MIDAS) regressions that use a combination of volatility measures estimates at different frequencies and horizons. Related, Engle and Gallo [137] exploit the information in different volatility measures, captured by a multivariate extension of the multiplicative error model suggested by Engle [136], to predict multi-step volatility. Finally, Andersen et al. [20] build on the Heterogeneous AutoRegressive (HAR) model by Barndorff-Nielsen and Shephard [50] and Corsi [110] and propose a HAR-RV component-based regression to forecast the  $h$ -steps ahead quadratic variation:

$$RV(t+h, h) = \beta_0 + \beta_D RV(t, 1) + \beta_W RV(t, 5) + \beta_M RV(t, 21) + \varepsilon(t+h). \quad (29)$$

Here the lagged volatility components  $RV(t, 1)$ ,  $RV(t, 5)$ , and  $RV(t, 21)$  combine to provide a parsimonious approximation to the long-memory type behavior of the realized volatility series, which has been documented in several studies (e. g., Andersen et al. [19]). Simple OLS esti-

mation yields consistent estimates for the coefficients in the regression (29), which can be used to forecast volatility out of sample.

As mentioned previously, the convergence results illustrated in Eqs. (27) and (28) stem from the theory of semi-martingales under conditions more general than those underlying the continuous-time diffusion in Eq. (22). For instance, these results are robust to the presence of discontinuities in the return path as in the jump-diffusion SV model (15). In this case the realized volatility measure (21) still converges to the return quadratic variation, which is now the sum of the diffusive integrated volatility IV and the cumulative squared jump component:

$$QV(t, h) = IV(t, h) + \sum_{t-h \leq s \leq t} J(s)^2. \quad (30)$$

The decomposition in Eq. (30) motivates the quest for separate estimates of the two quadratic variation components, IV and squared jumps. This is a fruitful exercise in forecasting applications, since separate estimation of the two components increases predictive accuracy (e. g., [20]). Further, this decomposition is relevant for derivatives pricing, e. g., options are highly sensitive to jumps as well as large moves in volatility (e. g., [141,220]).

A consistent estimate of integrated volatility is the  $k$ -skip bipower variation, BV (e. g., Barndorff-Nielsen and Shephard [53]),

$$BV(t, h; k, n) = \frac{\pi}{2} \sum_{i=k+1}^n \left| r \left( t-h + \frac{ih}{n}, \frac{h}{n} \right) \right| \times \left| r \left( t-h + \frac{(i-k)h}{n}, \frac{h}{n} \right) \right|. \quad (31)$$

Liu and Maheu [202] and Forsberg and Ghysels [147] show that realized power variation, which is robust to the presence of jumps, can improve volatility forecasts. A well-known special case of (31) is the ‘realized bipower variation’, which has  $k = 1$  and is denoted  $BV(t, h; n) \equiv BV(t, h; 1, n)$ . We can combine bipower variation with the realized volatility RV to obtain a consistent estimate of the squared jump component, i. e.,

$$RV(t, h; n) - BV(t, h; n) \xrightarrow{n \rightarrow \infty} QV(t, h) - IV(t, h) = \sum_{t-h \leq s \leq t} J(s)^2. \quad (32)$$

The result in Eq. (32) are useful to design tests for the presence of jumps in volatility, e. g., Andersen et al. [20], Barndorff-Nielsen and Shephard [53,54], Huang and

Tauchen [172], and Mizrach [217]. More recently, alternative approaches to test for jumps have been developed by Ait-Sahalia and Jacod [6], Andersen et al. [23], Lee and Mykland [195], and Zhang [261].

## Applications

The power of the continuous-time paradigm has been evident ever since the work by Merton [212] on intertemporal portfolio choice, Black and Scholes [63] on option pricing, and Vasicek [255] on bond valuation. However, the idea of casting these problems in a continuous-time diffusion context goes back all the way to the work in 1900 by Bachelier [32].

Merton [213] develops a continuous-time general-equilibrium intertemporal asset pricing model which is later extended by Cox et al. [112] to a production economy. Because of its flexibility and analytical tractability, the Cox et al. [112] framework has become a key tool used in several financial applications, including the valuation of options and other derivative securities, the modeling of the term structure of risk-free interest rates, the pricing of foreign currencies and defaultable bonds.

Volatility has played a central role in these applications. For instance, an option's payoff is non-linear in the price of the underlying asset and this feature renders the option value highly sensitive to the volatility of underlying returns. Further, derivatives markets have grown rapidly in size and complexity and financial institutions have been facing the challenge to manage intricate portfolios exposed to multiple risk sources. Risk management of these sophisticated positions hinges on volatility modeling. More recently, the markets have responded to the increasing hedging demands of investors by offering a menu of new products including, e.g., volatility swaps and derivatives on implied volatility indices like the VIX. These innovations have spurred an even more pressing need to accurately measure and forecast volatility in financial markets.

Research has responded to these market developments. We next provide a brief illustrative overview of the recent literature dealing with option pricing and term structure modeling, with an emphasis on the role that volatility modeling has played in these two key applications.

## Options

Rubinstein [233] and Bates [55], among others, note that prior to the 1987 market crash the Black and Scholes [63] (BS) formula priced option contracts quite accurately whereas after the crash it has been systematically underpricing out-of-the-money equity-index put contracts.

This feature is evident from Fig. 1, which is constructed from options on the S&P 500 futures. It shows the implied volatility function for near-maturity contracts traded both before and after October 19, 1987 ('Black Monday'). The mild u-shaped pattern prevailing in the pre-crash implied volatilities is labeled a 'volatility smile,' in contrast to the asymmetric post-1987 'volatility smirk'. Importantly, while the steepness and level of the implied volatility curve fluctuate day to day depending on market conditions, the curve has been asymmetric and upward sloping ever since 1987, so the smirk remains in place to the current date, e.g., Benzoni et al. [60]. In contrast, before the crash the implied volatility curve was invariably flat or mildly u-shaped as documented in, e.g., [57]. Finally, we note that the post-1987 asymmetric smirk for index options contrasts sharply with the pattern for individual equity options, which possess flat or mildly u-shaped implied volatility curves (e.g., [37,65]).

Given the failures of the BS formula, much research has gone into relaxing the underlying assumptions. A natural starting point is to allow volatility to evolve randomly, inspiring numerous studies that examine the option pricing implications of SV models. The list of early contributions includes [174,188,211,238,244,245,256]. Here we focus in particular on the Hull and White [174] model,

$$dp(t) = \mu_p dt + \sqrt{V(t)} dW_p(t) \quad (33)$$

$$\frac{dV(t)}{V(t)} = \mu_V dt + \sigma_V dW_V(t), \quad (34)$$

where  $W_p$  and  $W_V$  are standard Brownian motions. In general, shocks to returns and volatility may be (negatively) correlated, however for tractability Hull and White assume  $\rho = \text{corr}(dW_p, dW_V) = 0$ . Under this assumption they show that, in a risk-neutral world, the premium  $C^{\text{HW}}$  on a European call option is the Black and Scholes price  $C^{\text{BS}}$  evaluated at the average integrated variance  $\bar{V}$ ,

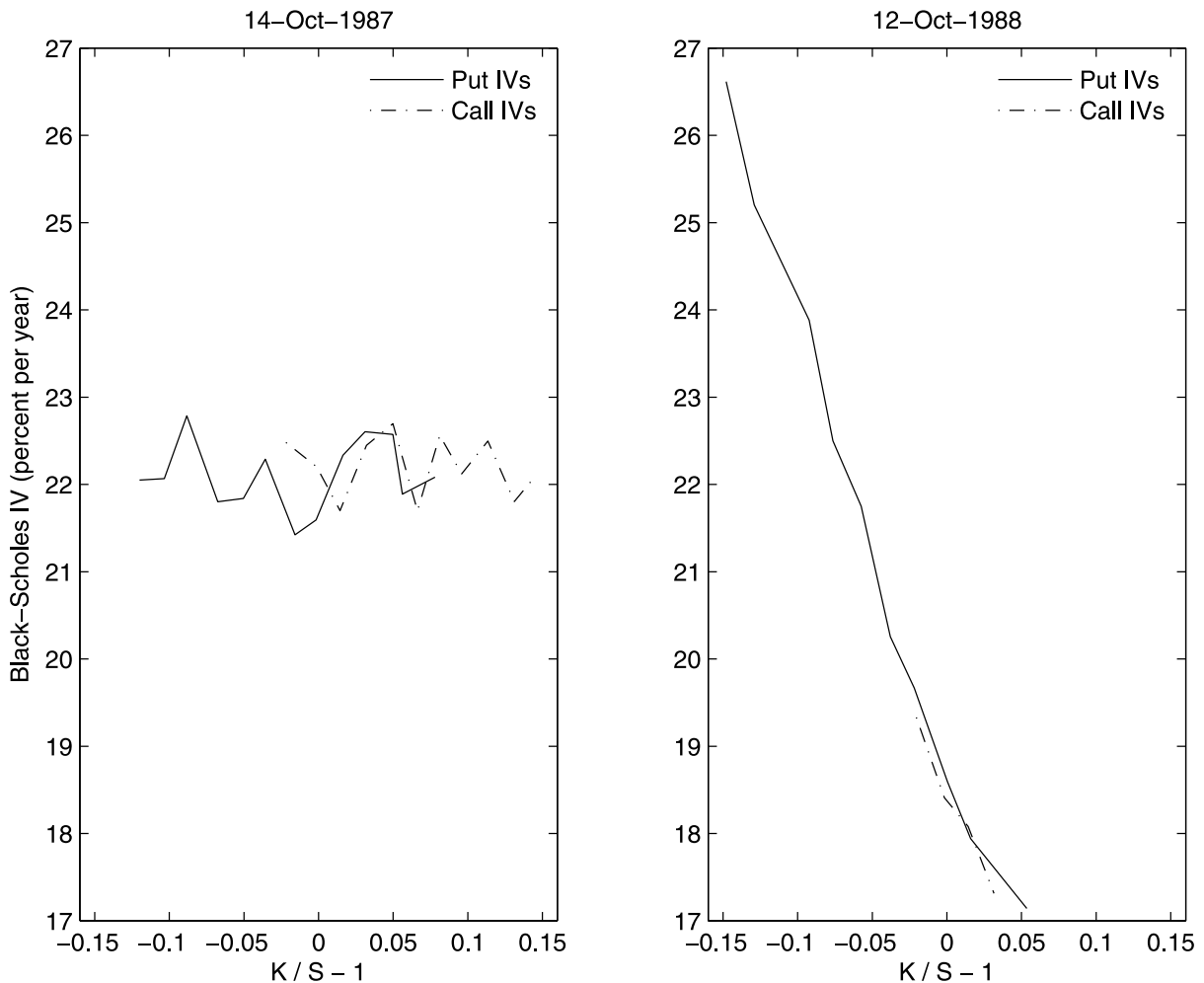
$$\bar{V} = \frac{1}{T-t} \int_t^T V(s) ds, \quad (35)$$

integrated over the distribution  $h(\bar{V}|V(t))$  of  $\bar{V}$ :

$$C^{\text{HW}}(p(t), V(t)) = \int C^{\text{BS}}(\bar{V}) h(\bar{V}|V(t)) d\bar{V}. \quad (36)$$

The early efforts to identify a more realistic probabilistic model for the underlying return were slowed by the analytical and computational complexity of the option pricing problem. Unlike the BS setting, the early SV specifications do not admit closed-form solutions. Thus, the evaluation of the option price requires time-consuming computations through, e.g., simulation methods or nu-





Stochastic Volatility, Figure 1

Pre- and post-1987 crash implied volatilities. The plots depict Black-Scholes implied volatilities computed from near-maturity options on the S&P 500 futures on October 14, 1987 (the week before the 1987 market crash) and a year later

merical solution of the pricing partial differential equation by finite difference methods. Further, the presence of a latent factor, volatility, and the lack of closed-form expressions for the likelihood function complicate the estimation problem.

Consequently, much effort has gone into developing restrictions for the distribution of the underlying return process that allow for (semi) closed-form solutions and are consistent with the empirical properties of the data. The ‘affine’ class of continuous-time models has proven particularly useful in providing a flexible, yet analytically tractable, setting. Roughly speaking, the defining feature of affine jump-diffusions is that the drift term, the conditional covariance term, and the jump intensity are all a linear-plus-constant (affine) function of the state vec-

tor. The Vasicek [255] bond valuation model and the Cox et al. [112] intertemporal asset pricing model provide powerful examples of the advantages of the affine paradigm.

To illustrate the progress in option pricing applications built on affine models, consider the return dynamics

$$dp(t) = \mu dt + \sqrt{V(t)}dW_p(t) + \xi_p(t)dq(t) \quad (37)$$

$$dV(t) = \kappa(\bar{V} - V(t))dt + \sigma_V \sqrt{V(t)}dW_V(t) + \xi_V(t)dq(t), \quad (38)$$

where  $W_p$  and  $W_V$  are standard Brownian motions with non-zero correlation  $\rho = \text{corr}(dW_p, dW_V)$ ,  $q$  is a Poisson process, uncorrelated with  $W_p$  and  $W_V$ , with jump inten-

sity

$$\lambda(t) = \lambda_0 + \lambda_1 V(t), \quad (39)$$

that is,  $\text{Prob}(dq_t = 1) = \lambda(t)dt$ . The jump amplitudes variables  $\xi_p$  and  $\xi_V$  have distributions

$$\xi_V(t) \rightsquigarrow \exp(\bar{\xi}_V) \quad (40)$$

$$\xi_p(t)|\xi_V(t) \rightsquigarrow N\left(\bar{\xi}_p + \rho_{\xi}\xi_V(t), \sigma_p^2\right). \quad (41)$$

Here volatility is not only stochastic but also subject to jumps which occur simultaneously with jumps in the underlying return process. The Black and Scholes model is a special case of (37)–(41) for constant volatility,  $V(t) = \sigma^2$ ,  $0 \leq t \leq T$ , and no jumps,  $\lambda(t) = 0$ ,  $0 \leq t \leq T$ . The Merton [214] model arises from (37)–(41) if volatility is constant but we allow for jumps in returns.

More recently, Heston [170] has considered a special case of (37)–(41) with stochastic volatility but without jumps. Using transform methods he derives a European option pricing formula which may be evaluated readily through simple numerical integration. His SV model has GARCH-type features, in that the variance is persistent and mean reverts at a rate  $\kappa$  to the long-run mean  $\bar{V}$ . Compared to Hull and White's [174] setting, Heston's model allows for shocks to returns and volatility to be negatively correlated, i. e.,  $\rho < 0$ , which creates a leverage-type effect and skews the return distribution. This feature is consistent with the properties of equity index returns. Further, a fatter left tail in the return distribution results in a higher cost for crash insurance and therefore makes out-of-the-money put options more expensive. This is qualitatively consistent with the patterns in implied volatilities observed after the 1987 market crash and discussed above.

Bates [56] has subsequently extended Heston's approach to allow for jumps in returns and using similar transform methods he has obtained a semi-closed form solution for the option price. The addition of jumps provides a more realistic description of equity returns and has important option pricing implications. With diffusive shocks (e. g., stochastic volatility) alone a large drop in the value of the underlying asset over a short time span is very unlikely whereas a market crash is always possible as long as large negative jumps can occur. This feature increases the value of a short-dated put option, which offers downside protection to a long position in the underlying asset.

Finally, Duffie et al. [130] have introduced a general model with jumps to volatility which embeds the dynamics (37)–(41). In model (37)–(41), the likelihood of a jump to occur increases when volatility is high ( $\lambda_1 > 0$ ) and a jump in returns is accompanied by an outburst of volatility. This

is consistent with what is typically observed during times of market stress. As in the Heston case, variance is persistent with a mean reversion coefficient  $\kappa$  towards its *diffusive* long-run mean  $\bar{V}$ , while the total long-run variance mean is the sum of the diffusive and jump components. In the special case of constant jump intensity, i. e.,  $\lambda_1 = 0$ , the total long-run mean is  $\bar{V} + \bar{\xi}_V \lambda_0 / \kappa$ . The jump term ( $\xi_V(t)dq(t)$ ) fattens the right tail of the variance distribution, which induces leptokurtosis in the return distribution. Two effects generate asymmetrically distributed returns. The first channel is the diffusive leverage effect, i. e.,  $\rho < 0$ , the second is the correlation between the volatility and the jump amplitude of returns generated through the coefficient  $\rho_{\xi}$ . Taken together, these effects increase model-implied option prices and help produce a realistic volatility smirk.

Several empirical studies rely on models of the form (37)–(41) in option-pricing applications. For instance, Bates [56] uses Deutsche Mark options to estimate a model with stochastic volatility and constant-intensity jumps to returns, while Bates [57] fits a jump-diffusion model with two SV factors to options on S&P 500 futures. In the latter case, the two SV factors combine to help capture features of the long-run memory in volatility while retaining the analytical tractability of the affine setting (see, e. g., [101] for another model with similar features). Alternative approaches to model long memory in continuous-time SV models rely on the fractional Brownian motion process, e. g., Comte and Renault [108] and Comte et al. [107], while Breidt et al. [76], Harvey [166] and Deo et al. [118] consider discrete-time SV models (see [175] for a review). Bakshi et al. [34,37] estimate a model similar to the one introduced by Bates [56] using S&P 500 options.

Other scholars rely on underlying asset return data alone for estimation. For instance, Andersen et al. [15] and Chernov et al. [95] use equity-index returns to estimate jump-diffusion SV models within and outside the affine (37)–(41) class. Eraker et al. [142] extend this analysis and fit a model that includes constant-intensity jumps to returns and volatility.

Finally, another stream of work examines the empirical implications of SV jump-diffusions using a joint sample of S&P 500 options and index returns. For example, Benzoni [59], Chernov and Ghysels [93], and Jones [190] estimate different flavors of the SV model without jumps. Pan [220] fits a model that has jumps in returns with time-varying intensity, while Eraker [141] extends Pan's work by adding jumps in volatility.

Overall, this literature has established that the SV jump-diffusion model dramatically improves the fit of underlying index returns and options prices compared to the

Black and Scholes model. Stochastic volatility alone has a first-order effect and jumps further enhance model performance by generating fatter tails in the return distribution and reducing the pricing error for short-dated options. The benefits of the SV setting are also significant in hedging applications.

Another aspect related to the specification of SV models concerns the pricing of volatility and jump risks. Stochastic volatility and jumps are sources of uncertainty. It is an empirical issue to determine whether investors demand to be compensated for bearing such risks and, if so, what the magnitude of the risk premium is. To examine this issue it is useful to write model (37)–(41) in so-called risk-neutral form. It is common to assume that the volatility risk premium is proportional to the instantaneous variance,  $\eta(t) = \eta_V V(t)$ . Further, the adjustment for jump risk is accomplished by assuming that the amplitude  $\tilde{\xi}_p(t)$  of jumps to returns has mean  $\bar{\tilde{\xi}}_p = \bar{\xi}_p + \eta_p$ . These specifications are consistent with an arbitrage-free economy. More general specifications can also be supported in a general equilibrium setting, e.g., a risk adjustment may apply to the jump intensity  $\lambda(t)$ . However, the coefficients associated to these risk adjustments are difficult to estimate and to facilitate identification they typically are fixed at zero. Incorporating such risk premia in model (37)–(41) yields the following risk-neutral return dynamics (e.g., Pan [220] and Eraker [141]):

$$dp(t) = (r - \mu^*)dt + \sqrt{V(t)}d\tilde{W}_p(t) + \tilde{\xi}_p(t)dq(t) \quad (42)$$

$$dV(t) = [\kappa(\bar{V} - V(t)) + \eta_V V(t)]dt + \sigma_V \sqrt{V(t)}d\tilde{W}_V(t) + \xi_V(t)dq(t), \quad (43)$$

where  $r$  is the risk-free rate,  $\mu^*$  a jump compensator term,  $\tilde{W}_p$  and  $\tilde{W}_V$  are standard Brownian motions under this so-called  $\mathcal{Q}$  measure, and the risk-adjusted jump amplitude variable  $\tilde{\xi}_p$  is assumed to follow the distribution,

$$\tilde{\xi}_p(t)|\xi_V(t) \rightsquigarrow N\left(\bar{\tilde{\xi}}_p + \rho_{\xi}\xi_V(t), \sigma_p^2\right). \quad (44)$$

Several studies estimate the risk-adjustment coefficients  $\eta_V$  and  $\eta_p$  for different specifications of model (37)–(44); see, e.g., Benzoni [59], Broadie et al. [78], Chernov and Ghysels [93], Eraker [141], Jones [190], and Pan [220]. It is found that investors demand compensation for volatility and jump risks and these risk premia are important for the pricing of index options. This evidence is reinforced by other studies examining the pricing of volatility risk using less structured but equally compelling procedures. For instance, Coval and Shumway [111] find that the returns

on zero-beta index option straddles (i.e., combinations of calls and puts that have offsetting covariances with the index) are significantly lower than the risk-free return. This evidence suggests that in addition to market risk at least a second factor (likely, volatility) is priced in the index option market. Similar conclusions are obtained by Bakshi and Kapadia [36], Buraschi and Jackwerth [79], and Broadie et al. [78].

### Risk-Free Bonds and Their Derivatives

The market for (essentially) risk-free Treasury bonds is liquid across a wide maturity spectrum. No-arbitrage restrictions constrain the allowable dynamics in the cross-section of bond yields. Much work has gone into the development of tractable dynamic term structure models capable of capturing the salient time-series properties of interest rates while respecting such cross-sectional no-arbitrage conditions. The class of so-called ‘affine’ dynamic term structure models provides a flexible and arbitrage-free, yet analytically tractable, setting for capturing the dynamics of the term structure of interest rates. Following Duffie and Kan [129], Dai and Singleton [114,115], and Piazzesi [226], the short term interest rate,  $y_0(t)$ , is an affine (i.e., linear-plus-constant) function of a vector of state variables,  $X(t) = \{x_i(t), i = 1, \dots, N\}$ :

$$y_0(t) = \delta_0 + \sum_{i=1}^N \delta_i x_i(t) = \delta_0 + \delta'_X X(t), \quad (45)$$

where the state-vector  $X$  has risk-neutral dynamics

$$dX(t) = \tilde{\mathcal{K}}(\tilde{\Theta} - X(t))dt + \Sigma \sqrt{S(t)}d\tilde{W}(t). \quad (46)$$

In Eq. (46),  $\tilde{W}$  is an  $N$ -dimensional Brownian motion under the so-called  $\mathcal{Q}$ -measure,  $\tilde{\mathcal{K}}$  and  $\tilde{\Theta}$  are  $N \times N$  matrices, and  $S(t)$  is a diagonal matrix with the  $i$ th diagonal element given by  $[S(t)]_{ii} = \alpha_i + \beta'_i X(t)$ . Within this setting, the time- $t$  price of a zero-coupon bond with time-to-maturity  $\tau$  is given by

$$P(t, \tau) = e^{A(\tau) - B(\tau)' X(t)}, \quad (47)$$

where the functions  $A(\tau)$  and  $B(\tau)$  solve a system of ordinary differential equations (ODEs); see, e.g., Duffie and Kan [129]. Semi-closed form solutions are also available for bond derivatives, e.g., bond options as well as caps and floors (e.g., Duffie et al. [130]).

In empirical applications it is important to also establish the evolution of the state vector  $X$  under the physical probability measure  $\mathcal{P}$ , which is linked to the  $\mathcal{Q}$ -dynamics (46) through a market price of risk,  $\Lambda(t)$ . Following Dai

and Singleton [114] the market price of risk is often given by

$$\Lambda(t) = \sqrt{S(t)}\lambda, \quad (48)$$

where  $\lambda$  is an  $N \times 1$  vector of constants. More recently, Duffee [127] proposed a broader ‘essentially affine’ class, which retains the tractability of standard models but, in contrast to the specification in Eq. (48), allows compensation for interest rate risk to vary independently of interest rate volatility. This additional flexibility proves useful in forecasting future yields. Subsequent generalization are in Duarte [124] and Cheridito et al. [92].

Litterman and Scheinkman [201] demonstrate that virtually all variation in US Treasury rates is captured by three factors, interpreted as changes in ‘level’, ‘steepness’, and ‘curvature’. Consistent with this evidence, much of the term-structure literature has focused on three-factor models. One problem with these models, however, is that the factors are latent variables void of immediate economic interpretation. As such, it is challenging to impose appropriate identifying conditions for the model coefficients and in particular to find the ideal representation for the ‘most flexible’ model, i. e., the model with the highest number of identifiable coefficients. Dai and Singleton [114] conduct an extensive specification analysis of multi-factor affine term structure models. They classify these models into subfamilies according to the number of (independent linear combination of) state variables that determine the conditional variance matrix of the state vector. Within each subfamily, they proceed to identify the models that lead to well-defined bond prices (a condition they label ‘admissibility’) and among the admissible specifications they identify a ‘maximal’ model that nests econometrically all others in the subfamily. Joslin [191] builds on Dai and Singleton’s [114] work by pursuing identification through a normalization of the drift term in the state vector dynamics (instead of the diffusion term, as in Dai and Singleton [114]). Duffie and Kan [129] follow an alternative approach to obtain an identifiable model by rotating from a set of latent state variables to a set of observable zero-coupon yields. Collin-Dufresne et al. [104] build on the insights of both Dai and Singleton [114] and Duffie and Kan [129]. They perform a rotation of the state vector into a vector that contains the first few components in the Taylor series expansion of the yield curve around a maturity of zero and their quadratic variation. One advantage is that the elements of the rotated state vector have an intuitive and unique economic interpretation (such as level, slope, and curvature of the yield curve) and therefore the model coefficients in this representation are identifiable. Further,

it is easy to construct a model-independent proxy for the rotated state vector, which facilitates model estimation as well as interpretation of the estimated coefficients across models and sample periods.

This discussion underscores an important feature of affine term structure models. The dependence of the conditional factor variance  $S(t)$  on one or more of the elements in  $X$  introduces stochastic volatility in the yields. However, when a square-root factor is present parametric restrictions (admissibility conditions) need to be imposed so that the conditional variance  $S(t)$  is positive over the range of  $X$ . These restrictions affect the correlations among the factors which, in turn, tend to worsen the cross-sectional fit of the model. Specifically, CIR models in which  $S(t)$  depends on all the elements of  $X$  require the conditional correlation among the factors to be zero, while the admissibility conditions imposed on the matrix  $\mathcal{K}$  renders the unconditional correlations non-negative. These restrictions are not supported by the data. In contrast, constant-volatility Gaussian models with no square-root factors do not restrict the signs and magnitude of the conditional and unconditional correlations among the factors but they do, of course, not accommodate the pronounced and persistent volatility fluctuations observed in bond yields. The class of models introduced by Dai and Singleton [114] falls between these two extremes. By including both Gaussian and square-root factors they allow for time-varying conditional volatilities of the state variables and yet they do not constrain the signs of some of their correlations. This flexibility helps to address the trade off between generating realistic correlations among the factors while capturing the time-series properties of the yields’ volatility.

A related aspect of (unconstrained) affine models concerns the dual role that square-root factors play in driving the time-series properties of yields’ volatility and the term structure of yields. Specifically, the time- $t$  yield  $y_\tau(t)$  on a zero-coupon bond with time-to-maturity  $\tau$  is given by

$$P(t, \tau) = e^{-\tau y_\tau(t)}. \quad (49)$$

Thus, we have

$$y_\tau(t) = -\frac{A(\tau)}{\tau} + \frac{B(\tau)'}{\tau} X(t). \quad (50)$$

It is typically assumed that the  $B$  matrix has full rank and therefore Eq. (50) provides a direct link between the state-vector  $X(t)$  and the term-structure of bond yields. Further, Itô’s Lemma implies that the yield  $y_\tau$  also follows a diffusion process:

$$dy_\tau(t) = \mu_{y_\tau}(X(t), t)dt + \frac{B(\tau)'}{\tau} \Sigma \sqrt{S(t)} d\tilde{W}(t). \quad (51)$$

Consequently, the (instantaneous) quadratic variation of the yield given as the squared yield volatility coefficient for  $y_\tau$  is

$$V_{y_\tau}(t) = \frac{B(\tau)'}{\tau} \Sigma S(t) \Sigma' \frac{B(\tau)}{\tau}. \quad (52)$$

The elements of the  $S(t)$  matrix are affine in the state vector  $X(t)$ , i. e.,  $[S(t)]_{ii} = \alpha_i + \beta_i' X(t)$ . Further, invoking the full rank condition on  $B(\tau)$ , Eq. (50) implies that each state variable in the vector  $X(t)$  is an affine function of the bond yields  $Y(t) = \{y_{\tau_j}(t), j = 1, \dots, J\}$ . Thus, for any  $\tau$  there is a set of constants  $a_{\tau,j}$ ,  $j = 0, \dots, J$ , so that

$$V_{y_\tau}(t) = a_{\tau,0} + \sum_{j=1}^J a_{\tau,j} y_{\tau_j}(t). \quad (53)$$

Hence, the current quadratic yield variation for bonds at any maturity is a linear combination of the term structure of yields. As such, the market is complete, i. e., volatility is perfectly spanned by a portfolio of bonds.

Collin-Dufresne and Goldstein [103] note that this spanning condition is unnecessarily restrictive and propose conditions which ensures that volatility no longer directly enters the main bond pricing Eq. (47). This restriction, which they term ‘unspanned stochastic volatility’ (USV), effectively breaks the link between the yields’ quadratic variation and the level of the term structure by imposing a reduced rank condition on the  $B(\tau)$  matrix. Further, since their model is a special (nested) case of the affine class it retains the analytical tractability of the affine model class. Recently Joslin [191] has derived more general conditions for affine term structure models to exhibit USV. His restrictions also produce a market incompleteness (i. e., volatility cannot be hedged using a portfolio of bonds) but do not constrain the degree of mean reversion of the other state variables so that his specification allows for more flexibility in capturing the persistence in interest rate series. (See also the USV conditions in the work by Trolle and Schwartz [253]).

There is conflicting evidence on the volatility spanning condition in fixed income markets. Collin-Dufresne and Goldstein [103] find that swap rates have limited explanatory power for returns on at-the-money ‘straddles’, i. e., portfolios mainly exposed to volatility risk. Similar findings are in Heidari and Wu [169], who show that the common factors in LIBOR and swap rates explain only a limited part of the variation in the swaption implied volatilities. Moreover, Li and Zhao [197] conclude that some of the most sophisticated multi-factor dynamic term structure models have serious difficulties in hedging caps and

cap straddles, even though they capture bond yields well. In contrast, Fan et al. [143] argue that swaptions and even swaption straddles can be well hedged with LIBOR bonds alone, supporting the notion that bond markets are complete.

More recently other studies have examined several versions of the USV restriction, again coming to different conclusions. A direct comparison of these results, however, is complicated by differences in the model specification, the estimation method, and the data and sample period used in the estimation. Collin-Dufresne et al. [105] consider swap rates data and fit the model using a Bayesian Markov Chain Monte Carlo method. They find that a standard three-factor model generates a time series for the variance state variable that is essentially unrelated to GARCH estimates of the quadratic variation of the spot rate process or to implied variances from options, while a four-factor USV model generates both realistic volatility estimates and a good cross-sectional fit. In contrast, Jacobs and Karoui [178] consider a longer data set of US Treasury yields and pursue quasi-maximum likelihood estimation. They find the correlation between model-implied and GARCH volatility estimates to be high. However, when estimating the model with a shorter sample of swap rates, they find such correlations to be small or negative. Thompson [250] explicitly tests the Collin-Dufresne and Goldstein [103] USV restriction and rejects it using swap rates data. Bikbov and Chernov [62], Han [164], Jarrow et al. [183], Joslin [192], and Trolle and Schwartz [254] rely on data sets of derivatives prices and underlying interest rates to better identify the volatility dynamics.

Andersen and Benzoni [12] directly relate model-free realized volatility measures (constructed from high-frequency US Treasury data) to the cross-section of contemporaneous bond yields. They find that the explanatory power of such regressions is very limited, which indicates that volatility is not spanned by a portfolio of bonds. The evidence in Andersen and Benzoni [12] is consistent with the USV models of Collin-Dufresne et al. [105] and Joslin [191], as well as with a model that embeds weak dependence between the yields and volatility as in Joslin [192]. Moreover, Duarte [125] argues that the effects of mortgage-backed security hedging activity affects both the interest rate volatility implied by options and the actual interest rate volatility. This evidence suggests that variables that are not in the span of the term structure of yields and forward rates contribute to explain volatility in fixed income markets. Also related, Wright and Zhou [258] find that adding a measure of market jump volatility risk to a regression of excess bond returns on the term structure of forward rates nearly doubles the  $R^2$  of the regression.



Taken together, these findings suggest more generally that genuine SV models are critical for appropriately capturing the dynamic evolution of the term structure.

### Estimation Methods

There are a very large number of alternative approaches to estimation and inference for parametric SV models and we abstain from a thorough review. Instead, we point to the basic challenges that exist for different types of specifications, how some of these were addressed in the early literature and finally provide examples of methods that have been used extensively in recent years. Our exposition continues to focus on applications to equity returns, interest rates, and associated derivatives.

Many of the original SV models were cast in discrete time, inspired by the popular GARCH paradigm. In that case, the distinct challenge for SV models is the presence of a strongly persistent latent state variable. However, more theoretically oriented models, focusing on derivatives applications, were often formulated in continuous time. Hence, it is natural that the econometrically-oriented literature has moved in this direction in recent years as well. This development provides an added complication as the continuous-time parameters must be estimated from discrete return data and without direct observations on volatility. For illustration, consider a fully parametric continuous-time SV model for the asset return  $r$  with conditional variance  $V$  and coefficient vector  $\Psi$ . Most methods to estimate  $\Psi$  rely on the conditional density  $f$  for the data generating process,

$$f(r(t), V(t)|\mathcal{I}(t-1), \Psi) = f_{r|V}(r(t)|V(t), \mathcal{I}(t-1), \Psi) \times f_V(V(t)|\mathcal{I}(t-1), \Psi), \quad (54)$$

where  $\mathcal{I}(t-1)$  is the available information set at time  $t-1$ . The main complications are readily identified. First, analytic expressions for the discrete-time transition (conditional) density,  $f$ , or the discrete-time moments implied by the data generating process operating in continuous time, are often unavailable. Second, volatility is latent in SV models, so that even if a closed-form expression for  $f$  is known, direct evaluation of the above expression is infeasible due to the absence of explicit volatility measures. The marginal likelihood with respect to the observable return process alone is obtained by integrating over all possible paths for the volatility process, but this integral has a dimension corresponding to sample size, rendering the approach infeasible in general.

Similar issues are present when estimating continuous-time dynamic term structure models. Following Pi-

azzesi [227], a change of variable gives the conditional density for a zero-coupon yield  $y$  on a bond with time to maturity  $\tau$ :

$$f(y_\tau(t)|\mathcal{I}(t-1), \Psi) = f_X(g(y_\tau(t), \Psi)|\mathcal{I}(t-1), \Psi) \times |\nabla_y g(y_\tau(t), \Psi)|. \quad (55)$$

Here the latent state vector  $X$  has conditional density  $f_X$ , the function  $g(\cdot, \Psi)$  maps the observable yield  $y$  into  $X$ ,  $X(t) = g(y_\tau(t), \Psi)$ , and  $\nabla_y g(y_\tau(t), \Psi)$  is the Jacobian determinant of the transformation. Unfortunately, analytic expressions for the conditional density  $f_X$  are known only in some special cases. Further, the mapping  $X(t) = g(y_\tau(t), \Psi)$  holds only if the model provides an exact fit to the yields, while in practice different sources of error (e.g., model mis-specification, microstructure effects, measurement errors) inject a considerable degree of noise into this otherwise deterministic linkage (for correct model specification) between the state vector and the yields. As such, a good measure of  $X$  might not be available to evaluate the conditional density (55).

### Estimation via Discrete-Time Model Specification or Approximation

The first empirical studies have estimated discrete-time SV models via a (Generalized) Method of Moments procedure by matching a number of theoretical and sample moments, e.g., Chan et al. [89], Ho et al. [171], Longstaff and Schwartz [204], and Melino and Turnbull [211]. These models were either explicitly cast in discrete time or were seen as approximate versions of the continuous-time process of interest. Similarly, several authors estimate diffusive affine dynamic term structure models by approximating the continuous-time dynamics with a discrete-time process. If the error terms are stipulated to be normally distributed, the transition density of the discretized process is multivariate normal and computation of unconditional moments then only requires knowledge of the first two moments of the state vector. This result facilitates quasi-maximum likelihood estimation. In evaluating the likelihood function, some studies suggest using closed-form expressions for the first two moments of the continuous-time process instead of the moments of the discretized process (e.g., Fisher and Gilles [145] and Duffee [127]), thus avoiding the associated discretization bias. This approach typically requires some knowledge of the state of the system which may be obtained, imperfectly, through matching the system, given the estimated parameter vector, to a set of observed zero-coupon yields to infer the state vector  $X$ . A modern alternative is to use the so-

called particle filter as an efficient filtering procedure for the unobserved state variables given the estimated parameter vector. We provide more detailed accounts of both of these procedures later in this section.

Finally, a number of authors develop a simulated maximum likelihood method that exploit the specific structure of the discrete-time SV model. Early examples are Danielsson and Richard [117] and Danielsson [116] who exploit the Accelerated Gaussian Importance Sampler for efficient Monte Carlo evaluation of the likelihood. Subsequent improvements were provided by Fridman and Harris [149] and Liesenfeld and Richard [200], with the latter relying on Efficient Importance Sampling (EIS). In a second step, EIS can also be used for filtering the latent volatility state vector. In general, these inference techniques provide quite impressive efficiency but the methodology is not always easy to generalize beyond the structure of the basic discrete-time SV asset return model. We discuss the general inference problem for continuous-time SV models for which the lack of a closed-form expression for the transition density is an additional complicating factor in a later section.

### Filtering the Latent State Variable Directly During Estimation

Some early studies focused on direct ways to extract estimates of the latent volatility state variable in discrete-time SV asset return models. The initial approach was based on quasi-maximum likelihood (QML) methods exploiting the Kalman filter. This method requires a transformation of the SV model to a linear state-space form. For instance, Harvey and Shephard [168] consider a version of the Taylor's [249] discrete-time SV model,

$$p(t) = p(t-1) + \beta + \sqrt{V(t)}\varepsilon(t) \quad (56)$$

$$\log(V(t)) = \alpha + \phi \log(V(t-1)) + \eta(t), \quad (57)$$

where  $p$  is the logarithmic price,  $\varepsilon$  is a zero-mean error term with unit variance, and  $\eta$  is an independently-distributed error term with zero mean and variance  $\sigma_\eta^2$ .

Define  $y(t) = p(t) - p(t-1) - \beta$ , square the observations in Eq. (56), and take logarithms to obtain the *measurement equation*,

$$\ell(t) = \omega + h(t) + \xi(t), \quad (58)$$

where  $\ell(t) \equiv \log y(t)^2$ ,  $h(t) \equiv \log(V(t))$ . Further,  $\xi$  is a zero-mean disturbance term given by  $\xi(t) = \log(\varepsilon(t)^2) - E[\log(\varepsilon(t)^2)]$ ,  $\omega = \log(\sigma^2) + E[\log(\varepsilon(t)^2)]$ , and  $\sigma$  is a scale constant which subsumes the effect of the drift term  $\alpha$

in Eq. (57). The autoregression (57) yields the *transition equation*,

$$h(t) = \phi h(t-1) + \eta(t), \quad (59)$$

Taken together, Eqs. (58) and (59) are the linear state-space transformation of the SV model (56)–(57). If the joint distribution of  $\varepsilon$  and  $\eta$  is symmetric, i. e.,  $f(\varepsilon, \eta) = f(-\varepsilon, -\eta)$ , then the disturbance terms in the state-space form are uncorrelated even if  $\eta$  and  $\varepsilon$  are not. A possible dependence between  $\varepsilon$  and  $\eta$  allows the model to pick up some of the asymmetric behavior often observed in stock returns. Projection of  $[h(t) - E_{t-1} h(t)]$  over  $[\ell(t) - E_{t-1} \ell(t)]$  yields the Kalman filter estimate of the latent (logarithmic) variance process:

$$\begin{aligned} E_t h(t) &= E_{t-1} h(t) \\ &+ \frac{E\{[h(t) - E_{t-1} h(t)] \times [\ell(t) - E_{t-1} \ell(t)]\}}{E\{[\ell(t) - E_{t-1} \ell(t)]^2\}} \\ &\times [\ell(t) - E_{t-1} \ell(t)], \end{aligned} \quad (60)$$

where the conditional expectations  $E_{t-1} \ell(t)$  and  $E_{t-1} h(t)$  are given by:

$$E_{t-1} \ell(t) = \omega + E_{t-1} h(t) \quad (61)$$

$$E_{t-1} h(t) = \phi E_{t-1} h(t-1). \quad (62)$$

To start the recursion (60)–(62), the initial value  $E_0 h(0)$  is fixed at the long-run mean  $\log(\bar{V})$ .

Harvey and Shephard [168] estimate the model coefficients via quasi-maximum likelihood, i. e. by treating the errors  $\xi$  and  $\eta$  as though they were normal and maximizing the prediction-error decomposition form of the likelihood function obtained via the Kalman filter. Inference is valid as long as the standard errors are appropriately adjusted. In their application they rely on daily returns on the value-weighted US market index over 1967–1987 and daily returns for 30 individual stocks over 1974–1983. Harvey et al. [167] pursue a similar approach to fit a multivariate SV model to a sample of four exchange rate series from 1981 to 1985. One major drawback of the Kalman filter approach is that the finite sample properties can be quite poor because the error term,  $\xi$ , is highly non-Gaussian, see, e.g., Andersen, Chung, and Sørensen [27]. The method may be extended to accommodate various generalizations including long memory persistence in volatility as detailed in Ghysels, Harvey, and Renault [158].

A related literature, often exploited in multivariate settings, specifies latent GARCH-style dynamics for a state vector which governs the systematic evolution of a higher dimensional set of asset returns. An early representative of these specifications is in Diebold and Nerlove [120], who exploit the Kalman filter for estimation, while Fiorentina et al. [144] provide a likelihood-based estimation procedure using MCMC techniques. We later review the MCMC approach and the associated filtering application, e.g. the ‘particle filter’, in some detail.

The state-space form is also useful to characterize the dynamics of interest rates. Following, e.g., Piazzesi [226], for a discrete-time dynamic term structure model the measurement and transition equations are

$$y_\tau(t) = -\frac{A(\tau)}{\tau} + \frac{B(\tau)'}{\tau} X(t) + \xi_\tau(t) \quad (63)$$

$$X(t) = \mu + \Phi X(t-1) + \Sigma \sqrt{S(t)} \varepsilon(t), \quad (64)$$

where  $S(t)$  is a matrix whose elements are affine functions of the state vector  $X$ , and  $A$  and  $B$  solve a system of difference equations. When all the yields are observed with error (i.e.,  $\xi_\tau \neq 0 \forall \tau$ ,  $0 \leq \tau \leq T$ ), QML estimation of the system (63)–(64) via the extended Kalman filter method yields an estimate of the coefficient vector. Applications of this approach for the US term structure data include Campbell and Viceira [81], Gong and Remolona [161], and Pennacchi [225]. The extended Kalman filter involves a linear approximation of the relation between the observed data and the state variables, and the associated approximation error will produce biased estimates. Christoffersen et al. [99] raise this concern and recommend the use of the so-called unscented Kalman filter for estimation of systems in which the relation between data and state variables is highly non-linear, e.g., options data.

### Methods Accommodating the Lack of a Closed-Form Transition Density

We have so far mostly discussed estimation techniques for models with either a known transition density or one that is approximated by a discrete-time system. However, the majority of empirically-relevant continuous-time models do not possess explicit transition densities and alternative approaches are necessary. This problem leads us naturally towards the large statistics and econometric literature on estimation of diffusions from discretely-observed data. The vast majority of these studies assume that all relevant variables are observed so the latent volatility or yield curve state variables, integral to SV models, are not ac-

counted for. Nonetheless, it may be feasible to extract the requisite estimates of the state variable by alternate means, thus restoring the feasibility, albeit not efficiency, of the basic approach. Since the literature is large and not directly geared towards genuine SV models, we focus on methods that have seen use in applications involving latent state variables.

A popular approach is to invert the map between the state vector and a subset of the observables assuming that the model prices specific securities exactly. In applications to equity markets this is done, e.g., by assuming that one option contract is priced without error, which implies a specific value (estimate) of the variance process given the model parameters  $\Psi$ . For instance, Pan [220] follows this approach in her study of S&P 500 options and returns, which we review in more detail in Sect. “[Estimation from Option Data](#)”. In applications to fixed income markets it is likewise stipulated that certain bonds are priced without error, i.e., in Eq. (63) the error term  $\xi_{\tau_i}(t)$  is fixed at zero for a set of maturities  $\tau_1, \dots, \tau_N$ , where  $N$  matches the dimension of the state vector  $X$ . This approach yields an estimate for the latent variables through the inverse-map  $X(t) = g(y_\tau(t), \Psi)$ .

One criticism of the state vector inversion procedure is that it requires ad hoc assumptions regarding the choice of the securities that are error-free (those used to compute model-implied measures of the state vector) vis-a-vis those observed with error (used either for estimation or to assess model performance in an ‘out-of-sample’ cross-sectional check). In fact, the extracted state vector can be quite sensitive to the choice of derivatives (or yields) used. Nevertheless, this approach has intuitive appeal. Model-implied measures of the state vector, in combination with a closed-form expression for the conditional density (55), allow for efficient estimation of the coefficient vector  $\Psi$  via maximum likelihood. Analytic expressions for  $f_X$  in Eq. (55) exist in a limited number of cases. For instance, if  $X$  is Gaussian then  $f_X$  is multivariate normal, while if  $X$  follows a square-root process then  $f_X$  can be expressed in terms of the modified Bessel function (e.g., [113]). Different flavors of these continuous-time models are estimated in, e.g., [91,106,132,182,223]. In more general cases, including affine processes that combine Gaussian and square-root state variables, closed-form expressions for  $f_X$  are no longer available. In the rest of this section we briefly review different methods to overcome this problem. The interested reader may consult, e.g., [226] for more details.

Lo [203] warns that the common approach of estimating parameters of an Itô process by applying maximum likelihood to a discretization of the stochastic differen-

tial equation yields inconsistent estimators. In contrast, he characterizes the likelihood function as a solution to a partial differential equation. The method is very general, e. g., it applies not only to continuous-time diffusions but also to jump processes. In practice, however, analytic solutions to the partial differential equations (via, e. g., Fourier transforms) are available only for a small class of models so computationally-intensive methods (e. g., finite differencing or simulations) are generally required to solve the problem. This is a severe limitation in the case of multivariate systems like SV models.

For general Markov processes, where the above solution is infeasible, a variety of procedures have been advocated in recent years. Three excellent surveys provide different perspectives on the issue. Aït-Sahalia, Hansen, and Scheinkman [5] discuss operator methods and mention the potential of applying a time deformation technique to account for genuine SV features of the process, as in Conley, Hansen, Luttmer, and Scheinkman [109]. In addition, the Aït-Sahalia [3,4] closed-form polynomial expansions for discretely-sampled diffusions are reviewed along with the Schaumburg [235] extension to a general class of Markov processes with Lévy-type generators. Meanwhile, Bibby, Jacobsen, and Sørensen [61] survey the extensive statistics literature on estimating functions for diffusion-type models and Bandi and Phillips [42] explicitly consider dealing with nonstationary processes (see also the work of Bandi [39], Bandi and Nguyen [41], and Bandi and Phillips [43,44]).

The characteristic function based inference technique has been particularly widely adopted due to the natural fit with the exponentially affine model class which provides essentially closed-form solutions for many pricing applications. Consequently, we dedicate a separate section to this approach.

**Characteristic Functions** Singleton [242] proposes to exploit the information contained in the conditional characteristic function of the state vector  $X$ ,

$$\phi(iu, X(t), \Psi) = E[e^{iu'X(t+1)} | X(t)], \quad (65)$$

to pursue maximum likelihood estimation of affine term structure models. In Equation (65) we highlight the dependence of the characteristic function on the unknown parameter vector  $\Psi$ . When  $X$  is an affine (jump-)diffusion process,  $\phi$  has the exponential affine form,

$$\phi(iu, X(t), \Psi) = e^{\alpha_t(u) + \beta_t(u)'X(t)}, \quad (66)$$

where the functions  $\alpha$  and  $\beta$  solve a system of ODEs. As such, the transition density  $f_X$  is known explicitly up to an inverse-Fourier transformation of the characteristic function (65),

$$f_X(X(t+1) | X(t); \Psi) = \frac{1}{\pi^N} \int_{\mathbb{R}_+^N} \text{Re}[e^{-iu'X(t+1)} \phi(iu, X(t), \Psi)] du. \quad (67)$$

Singleton shows that Gauss–Legendre quadrature with a relatively small number of quadrature points allows to accurately evaluate the integral in Eq. (67) when  $X$  is univariate. As such, the method readily delivers efficient estimates of the parameter vector,  $\Psi$ , subject to an auxiliary assumption, namely that the state vector may be extracted by assuming that a pre-specified set of security prices is observed without error while the remainder have non-trivial error terms.

When  $X$  is multivariate the Fourier inversion in Eq. (67) is computationally more demanding. Thus, when estimating multi-dimensional systems Singleton suggests focusing on the conditional density function of the individual elements of  $X$ , but conditioned on the full state vector,

$$f_{X_j}(X_j(t+1) | X(t); \Psi) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-i\omega \mathbf{I}_j' X(t+1)} \phi(i\omega \mathbf{I}_j, X(t), \Psi) d\omega, \quad (68)$$

where the vector  $\mathbf{I}_j$  has 1 in the  $j$ th element and zero elsewhere so that the  $j$ th element of  $X$  is  $X_j(t+1) = \mathbf{I}_j' X(t+1)$ . Maximization of the likelihood function obtained from  $f_{X_j}$ , for a fixed  $j$ , will often suffice to obtain a consistent estimate of  $\Psi$ . Exploiting more than one of the conditional densities (68) will result in more efficient  $\Psi$  estimate. For instance, the scores of multiple univariate log-likelihood functions, stacked in a vector, yield moment conditions that allow for generalized method of moment (GMM) estimation of the system. Alternatively, Joslin [192] proposes a change-of-measure transformation which reduces the oscillatory behavior of the integrand in Eq. (67). When using this transformation, Gauss–Hermite quadrature more readily provides a solution to the integral in (67) even if the state vector  $X$  is multi-dimensional, thus facilitating full ML estimation of the system.

Related, several studies have pursued GMM estimation of affine processes using characteristic functions. Definition (65) yields the moment condition

$$E[(\phi(iu, X(t), \Psi) - e^{iu'X(t+1)})z(u, X(t))] = 0, \quad (69)$$

where  $X$  is an  $N$ -dimensional (jump-)diffusion,  $u \in \mathbb{R}^N$ , and  $z$  is an instrument function. When  $X$  is affine, the characteristic function takes the exponential form (66). Different choices of  $u$  and  $z$  yield a set of moment conditions that can be used for GMM estimation and inference. Singleton [242] derives the optimal instrument in terms of delivering efficient estimates. Carrasco et al. [86] approximate the optimal instrument with a set of basis functions that do not require the knowledge of the conditional likelihood function, thus avoiding one of the assumptions invoked by Singleton. Further, they build on Carrasco and Florens [87] to implement estimation using a continuum of moment conditions, which yields maximum-likelihood efficiency. Other applications of GMM-characteristic function methods to affine (jump-) diffusions for equity index returns are in Chacko and Viceira [88] and Jiang and Knight [184].

In some cases the lack of closed-form expressions for the moment condition in Eq. (69) can hinder GMM estimation. In these cases the expectation in Eq. (69) can be evaluated by Monte Carlo integration. This is accomplished by simulating a long sample from the discretized process for a given value of the coefficient vector  $\Psi$ . The parameter  $\Psi$  is then estimated via the simulated method of moments (SMM) of McFadden [206] and Duffie and Singleton [131]. Singleton [242] proposes SMM characteristic function estimators that exploit the special structure of affine term structure models.

### Efficient Estimation of General Continuous-Time Processes

A number of recent approaches offer excellent flexibility in terms of avoiding approximations to the continuous-time model-implied transition density while still facilitating efficient estimation of the evolution of the latent state vector for the system.

### Maximum Likelihood with Characteristic Functions

Bates [58] develops a filtration-based maximum likelihood estimation method for affine processes. His approach relies on Bayes' rule to recursively update the joint characteristic function of latent variables and data conditional on past data. He then obtains the transition density by Fourier inversion of the updated characteristic function.

Denote with  $y(t)$  and  $X(t)$  the time- $t$  values of the observable variable and the state vector, respectively, and let  $Y(t) \equiv \{y(1), \dots, y(t)\}$  be the data observed up to time  $t$ . Consider the case in which the characteristic function of  $z(t+1) \equiv (y(t+1), X(t+1))$  conditional on  $z(t) \equiv (y(t), X(t))$ , is an exponential affine function of

$X(t)$ :

$$\begin{aligned} \phi(is, iu, z(t), \Psi) &= \mathbb{E} \left[ e^{is'y(t+1) + iu'X(t+1)} | z(t) \right] \\ &= e^{\alpha(is, iu, y(t)) + \beta(is, iu, y(t))'X(t)}. \end{aligned} \quad (70)$$

Next, determine the value of the characteristic function conditional on the observed data  $Y(t)$ :

$$\begin{aligned} \phi(is, iu, Y(t), \Psi) &= \mathbb{E} \left[ \mathbb{E} \left[ e^{is'y(t+1) + iu'X(t+1)} | z(t) \right] | Y(t) \right] \\ &= \mathbb{E} \left[ e^{\alpha(is, iu, y(t)) + \beta(is, iu, y(t))'X(t)} | Y(t) \right] \\ &= e^{\alpha(is, iu, y(t))} \psi(\beta(is, iu, y(t)), Y(t), \Psi), \end{aligned} \quad (71)$$

where  $\psi(iu, Y(t), \Psi) \equiv \mathbb{E} \left[ e^{iu'X(t)} | Y(t) \right]$  denotes the (marginal) characteristic function for the state vector conditional on the observed data. Fourier inversion then yields the conditional density for the observation  $y(t+1)$  conditional on  $Y(t)$ :

$$\begin{aligned} f_y(y(t+1) | Y(t); \Psi) &= \frac{1}{2\pi} \int_{\mathbb{R}} e^{-is'y(t+1)} \phi(is, 0, Y(t), \Psi) ds. \end{aligned} \quad (72)$$

The next step updates the characteristic function  $\psi$  (Bartlett [48]):

$$\begin{aligned} \psi(iu, Y(t+1), \Psi) &= \frac{1}{2\pi f_y(y(t+1) | Y(t); \Psi)} \\ &\cdot \int_{\mathbb{R}} e^{-is'y(t+1)} \phi(is, iu, Y(t), \Psi) ds. \end{aligned} \quad (73)$$

To start the recursion, Bates initializes  $\psi$  at the unconditional characteristic function of the latent variable  $X$ . The log-likelihood function is then given by

$$\begin{aligned} \log \mathcal{L}(Y(T); \Psi) &= \log(f_y(y(1); \Psi)) \\ &+ \sum_{t=2}^T \log(f_y(y(t) | Y(t-1); \Psi)). \end{aligned} \quad (74)$$

A nice feature is that the method provides a natural solution to the filtering problem. The filtered estimate of the latent state  $X$  and its variance are computed from the first and second derivatives of the moment generating function  $\psi(u, Y(t); \Psi)$  in Eq. (73), evaluated at  $u = 0$ :

$$\begin{aligned} \mathbb{E}[X(t+1) | Y(t+1); \Psi] &= \frac{1}{2\pi f_y(y(t+1) | Y(t); \Psi)} \\ &\times \int_{\mathbb{R}} e^{-is'y(t+1)} \phi_u(is, 0, Y(t); \Psi) ds \end{aligned} \quad (75)$$



$$\begin{aligned} \text{Var}[X(t+1)|Y(t+1); \Psi] &= \frac{1}{2\pi f_y(y(t+1)|Y(t); \Psi)} \\ &\times \int_{\mathbb{R}} e^{-is'y(t+1)} \phi_{uu}(is, 0, Y(t); \Psi) ds \\ &- \{E[X(t+1)|Y(t+1)]\}^2. \quad (76) \end{aligned}$$

A drawback is that at each step  $t$  of the iteration the method requires storage of the entire characteristic function  $\psi(iu, Y(t); \Psi)$ . To deal with this issue Bates recommends to approximate the true  $\psi$  with the characteristic function of a variable with a two-parameter distribution. The choice of the distribution depends on the  $X$ -dynamics while the two parameters of the distribution are determined by the conditional mean  $E[X(t+1)|Y(t+1); \Psi]$  and variance  $\text{Var}[X(t+1)|Y(t+1); \Psi]$  given in Equations (75)–(76).

In his application Bates finds that the method is successful in estimating different flavors of the SV jump-diffusion for a univariate series of daily 1953–1996 S&P 500 returns. In particular, he shows that the method obtains estimates that are equally, if not more, efficient compared to the efficient method of moments and Markov Chain Monte Carlo methods described below. Extensions of the method to multivariate processes are theoretically possible, but they require numerical integration of multi-dimensional functions, which is computationally demanding.

**Simulated Maximum Likelihood** In Sect. “[Filtering the Latent State Variable Directly During Estimation](#)” we discussed methods for simulated ML estimation and inference in discrete-time SV models. Pedersen [224] and Santa-Clara [234] independently develop a simulated maximum likelihood (SML) method to estimate continuous-time diffusion models. They divide each interval in between two consecutive data points  $X_{t+1}$  and  $X_t$  into  $M$  sub-intervals of length  $\Delta = 1/M$  and they discretize the  $X$  process using the Euler scheme,

$$\begin{aligned} X_{t+(i+1)\Delta} &= X_{t+i\Delta} + \mu(X_{t+i\Delta})\Delta \\ &+ \Sigma(X_{t+i\Delta})\sqrt{\Delta}\varepsilon_{t+(i+1)\Delta}, \\ i &= 0, \dots, M-1, \quad (77) \end{aligned}$$

where  $\mu$  and  $\Sigma$  are the drift and diffusion terms of the  $X$  process and  $\varepsilon$  is multivariate normal with mean zero and identity variance matrix. The transition density of the discretized process is multivariate normal with mean  $\mu$  and variance matrix  $\Sigma \Delta$ . As  $\Delta$  goes to zero, this density converges to that of the continuous-time process  $X$ . As such,

the transition density from  $X_t$  to  $X_{t+1}$  is given by

$$\begin{aligned} f_X(X_{t+1}|X_t; \Psi) &= \int f_X(X_{t+1}|X_{t+1-\Delta}; \Psi) \\ &\times f_X(X_{t+1-\Delta}|X_t; \Psi) dX_{t+1-\Delta}. \quad (78) \end{aligned}$$

For sufficiently small values of  $\Delta$  the first term in the integrand,  $f_X(X_{t+1}|X_{t+1-\Delta}; \Psi)$ , is approximated by the transition density of the discretized process, while the second term,  $f_X(X_{t+1-\Delta}|X_t; \Psi)$ , is a multi-step-ahead transition density that can be computed from the recursion from  $X_t$  to  $X_{t+1-\Delta}$ . Writing the right-hand side of Eq. (78) as a conditional expectation yields

$$f_X(X_{t+1}|X_t; \Psi) = E_{X_{t+1-\Delta}|X_t}[f_X(X_{t+1}|X_{t+1-\Delta}; \Psi)]. \quad (79)$$

The expectation in Eq. (79) can be computed by Monte Carlo integration over a large number of paths for the process  $X$ , simulated via the Euler scheme (77). As  $\Delta$  vanishes, the Euler scheme is consistent. Thus, when the size of the simulated sample increases the sample average of the function  $f_X$ , evaluated at the random draws of  $X_{t+1-\Delta}$ , converges to the true transition density. Application of the principles in Bladt and Sørensen [64] may well be useful in enhancing the efficiency of the simulation scheme and hence the actual efficiency of the inference procedure in practice.

Brandt and Santa-Clara [75] apply the SML method to estimate a continuous-time model of the joint dynamics of interest rates in two countries and the exchange rate between the two currencies. Piazzesi [227] extends the SML approach for jump-diffusion processes with time-varying jump intensity. She considers a high-frequency policy rule based on yield curve information and an arbitrage-free bond market and estimates the model using 1994–1998 data on the Federal Reserve target rate, the six-month LIBOR rate, and swap yields.

An important issue is how to initialize any unobserved component of the state vector,  $X(t)$ , such as the volatility state at each observation to provide a starting point for the next Monte Carlo integration step. This may be remedied through application of the particle filter, as mentioned earlier and discussed below in connection with MCMC estimation. Another possibility is, as also indicated previously, to extract the state variable through inversion from derivatives prices or yields assumed observed without pricing errors.

**Indirect Inference** There are also other method-of-moments strategies to estimate finitely-sampled continuous-

time processes of a general type. One prominent approach approximates the unknown transition density for the continuous-time process with the density of a semi-nonparametric (SNP) auxiliary model. Then one can use the score function of the auxiliary model to form moment conditions for the parameter vector  $\Psi$  of the continuous-time model. This approach yields the efficient method of moments estimator (EMM) of Gallant and Tauchen [154], Gallant et al. [150], and Gallant and Long [152], and the indirect inference estimator of Gouriéroux et al. [162] and Smith [243].

To fix ideas, suppose that the conditional density for a continuous-time return process  $r$  (the ‘structural’ model) is unknown. We intend to approximate the unknown density with a discrete-time model (the ‘auxiliary’ model) that is tractable and yet sufficiently flexible to accommodate the systematic features of the actual data sample well. A parsimonious auxiliary density for  $r$  embeds ARMA and EGARCH leading terms to capture the conditional mean and variance dynamics. There may be residual excess skewness and kurtosis that elude the ARMA and EGARCH forms. As such, the auxiliary density is rescaled using a nonparametric polynomial expansion of order  $K$ , which yields

$$g_K(r(t)|x(t); \xi) = \left( \nu + (1 - \nu) \times \frac{[P_K(z(t), x(t))]^2}{\int_{\mathbb{R}} [P_K(z(t), x(t))]^2 \phi(u) du} \right) \frac{\phi(z(t))}{\sqrt{h(t)}}, \quad (80)$$

where  $\nu$  is a small constant,  $\phi(\cdot)$  is the standard normal density,  $x(t)$  contains lagged return observations, and

$$z(t) = \frac{r(t) - \mu(t)}{\sqrt{h(t)}}, \quad (81)$$

$$\mu(t) = \phi_0 + ch(t) + \sum_{i=1}^s \phi_i r(t-1) + \sum_{i=1}^u \delta_i \varepsilon(t-1), \quad (82)$$

$$\begin{aligned} \log h(t) = & \omega + \sum_{i=1}^p \beta_i \log h(t-1) \\ & + (1 + \alpha_1 L + \dots + \alpha_q L^q) \\ & \times \left[ \theta_1 z(t-1) + \theta_2 (b(z(t-1)) - \sqrt{2/\pi}) \right], \end{aligned} \quad (83)$$

$$P_K(z, x) = \sum_{i=0}^{K_z} a_i(x) z^i = \sum_{i=0}^{K_z} \left( \sum_{|j|=0}^{K_x} a_{ij} x^j \right) z^i, \quad (84)$$

$$a_{00} = 1.$$

Here  $j$  is a multi-index vector,  $x^j \equiv (x_1^{j_1}, \dots, x_M^{j_M})$ , and  $|j| \equiv \sum_{m=1}^M j_m$ . The term  $b(z)$  is a smooth (twice-differentiable) function that closely approximates the absolute value operator in the EGARCH variance equation.

In practice, the representation of  $P_K$  is given by Hermite orthogonal polynomials. When the order  $K$  of the expansion increases, the auxiliary density will approximate the data arbitrarily well. If the structural model is indeed the true data generating process, then the auxiliary density will converge to that of the structural model. For a given  $K$ , the QML estimator  $\hat{\xi}$  for the auxiliary model coefficient satisfies the score condition

$$\frac{1}{T} \sum_{t=1}^T \frac{\partial \log g_K(r(t)|x(t); \hat{\xi})}{\partial \xi} = 0. \quad (85)$$

Suppose now that the structural model is correct and  $\Psi_0$  is the true value of its coefficient vector. Consider a series  $\{r(t; \Psi), x(t; \Psi)\}$ ,  $t = 1, \dots, T(T)$ , simulated from the structural model. Then we expect that the score condition (85) holds when evaluated by averaging over the simulated returns rather than over the actual data:

$$\begin{aligned} m_{T(T)}(\Psi_0, \hat{\xi}) &= \frac{1}{T(T)} \sum_{t=1}^{T(T)} \frac{\partial \log g_K(r(t, \Psi_0)|x(t, \Psi_0); \hat{\xi})}{\partial \xi} \\ &\approx 0. \end{aligned} \quad (86)$$

When  $T$  and  $T(T)$  tend to infinity, condition (86) holds exactly.

Gallant and Tauchen [154] propose the EMM estimator  $\hat{\Psi}$  defined via

$$\hat{\Psi} = \arg \min_{\Psi} m_{T(T)}(\Psi, \hat{\xi})' \hat{W}_T m_{T(T)}(\Psi, \hat{\xi}), \quad (87)$$

where the weighting matrix  $\hat{W}_T$  is a consistent estimate of the inverse asymptotic covariance matrix of the auxiliary score function, e.g., the inverse outer product of the SNP gradient:

$$\begin{aligned} \hat{W}_T^{-1} = & \frac{1}{T} \sum_{t=1}^T \left[ \frac{\partial \log g_K(r(t)|x(t); \hat{\xi})}{\partial \xi} \right] \\ & \times \left[ \frac{\partial \log g_K(r(t)|x(t); \hat{\xi})}{\partial \xi} \right]'. \end{aligned} \quad (88)$$

An important advantage of the technique is that EMM estimates achieve the same degree of efficiency as the ML procedure, when the score of the auxiliary model asymptotically spans the score of the true model. It also delivers powerful specification diagnostics that provide guidance in the model selection. Gallant and Tauchen [154] show that the EMM estimator is asymptotically normal. Further, under the assumption that the structural model is correctly specified, they derive a  $\chi^2$  statistic for the test of over-identifying restrictions. Gallant et al. [150] normalize the vector  $m_{\mathcal{T}(T)}(\hat{\Psi}, \hat{\xi})$  by its standard error to obtain a vector of score  $t$ -ratios. The significance of the individual score elements is often informative of the source of model mis-specification, with the usual caveat that failure to capture one characteristic of the data may result in the significance of a moment condition that pertains to a coefficient not directly related to that characteristic (due to correlation in the moment conditions). Finally, EMM provides a straightforward solution to the problem of filtering and forecasting the latent return variance process  $V$ , i.e., determining the conditional densities  $f(V(t)|x(t), \Psi)$  and  $f(V(t+j)|x(t), \Psi)$ ,  $j \geq 0$ . This is accomplished through the *reprojection* method discussed in, e.g., Gallant and Long [152] and Gallant and Tauchen [155]. In applications to dynamic term structure models, the same method yields filtered and forecasted values for the latent state variables.

The reprojection method assumes that the coefficient vector  $\Psi$  is known. In practice,  $\Psi$  is fixed at the EMM estimate  $\hat{\Psi}$ . Then one simulates a sample of returns and latent variables from the structural model and fits the auxiliary model on the simulated data. This is equivalent to the first step of the EMM procedure except that, in the reprojection step, we fit the auxiliary model assuming the structural model is correct, rather than using actual data. The conditional density of the auxiliary model, estimated under the null, approximates the unknown density of the structural model:

$$g_K(r(t+j)|x(t); \hat{\xi}) \approx f(r(t+j)|x(t); \hat{\Psi}), \quad j \geq 0, \quad (89)$$

where  $\hat{\xi}$  is the QML estimate of the auxiliary model coefficients obtained by fitting the model on simulated data. This approach yields filtered estimates and forecasts for the conditional mean and variance of the return via

$$E[r(t+j)|x(t); \hat{\Psi}] = \int y g_K(y|x(t); \hat{\xi}) dy, \quad (90)$$

$$\begin{aligned} \text{Var}[r(t+j)|x(t); \hat{\Psi}] &= \int (y - E[r(t+j)|x(t); \hat{\Psi}])^2 \\ &\times g_K(y|x(t); \hat{\xi}) dy. \quad (91) \end{aligned}$$

An alternative approach consists in fitting an auxiliary model for the latent variable (e.g., the return conditional variance) as a function of current and lagged returns. It is straightforward to estimate such model using data on the latent variable and the associated returns simulated from the structural model with the EMM coefficient  $\hat{\Psi}$ . Also in this case the auxiliary model density approximates the true one, i.e.,

$$g_K^V(V(t+j)|x(t); \hat{\xi}) \approx f^V(V(t+j)|x(t); \hat{\Psi}), \quad j \geq 0. \quad (92)$$

This approach yields a forecast for the conditional variance process,

$$E[V(t+j)|x(t); \hat{\Psi}] = \int v g_K^V(v|x(t); \hat{\xi}) dv. \quad (93)$$

In sum, reprojection is a simulation approach to implement a non-linear Kalman-filter-type technique, which yields effective forecasts for the unobservable state vector.

The indirect inference estimator by Gouriéroux et al. [162] and Smith [243] is closely related to the EMM estimator. Indirect inference exploits that the following two quantities should be close when the structural model is correct and the data are simulated at the true parameter  $\Psi_0$ : (i) the QML estimator  $\hat{\xi}$  for the auxiliary model computed from actual data; (ii) the QML estimator  $\hat{\xi}(\Psi)$  for the auxiliary model fitted on simulations from the structural model. Minimizing the distance between  $\hat{\xi}$  and  $\hat{\xi}(\Psi)$  in an appropriate metric yields the indirect inference estimator for  $\Psi$ . Similar to EMM, asymptotic normality holds and a  $\chi^2$  test for over-identifying restrictions is available. However, the indirect inference approach is computationally more demanding, because finding the value of  $\Psi$  that minimizes the distance function requires re-estimating the auxiliary model on a different simulated sample for each iteration of the optimization routine. EMM does not have this drawback, since the EMM objective function is evaluated at the same fitted score at each iteration. Nonetheless, there may well be circumstances where particular auxiliary models are of primary economic interest and estimation based on the corresponding moment conditions may serve as a useful diagnostic tool for model performance in such directions.

Several studies have used EMM to fit continuous-time SV jump-diffusion models for equity index returns, e.g., Andersen et al. [15], Benzoni [59], Chernov and Ghysels [93], and Chernov et al. [94,95]. Andersen and Lund [28] and Andersen et al. [16] use EMM to estimate SV jump-diffusion models for the short-term inter-

est rate. Ahn et al. [1,2], Brandt and Chapman [71], and Dai and Singleton [114] fit different flavors of multi-factor dynamic term structure models. Andersen et al. [27] document the small-sample properties of the efficient method of moments estimator for stationary processes, while Duffee and Stanton [128] study its properties for near unit-root processes. A. Ronald Gallant and George E. Tauchen at Duke University have prepared well-documented general-purpose EMM and SNP packages, available for download at the web address [ftp.econ.duke.edu](http://ftp.econ.duke.edu) in the directories `pub/get/emm` and `pub/get/snp`. In applications it is often useful to customize the SNP density to allow for a more parsimonious fit of the data under investigation. For instance, Andersen et al. [15,16], Andersen and Lund [28], and Benzoni [59] rely on the SNP density (80)–(84).

**Markov Chain Monte Carlo** The MCMC method provides a Bayesian solution to the inference problem for a dynamic asset pricing model. The approach treats the model coefficient  $\Psi$  as well as the vector of latent state variables  $X$  as random variables and computes the posterior distribution  $f(\Psi, X|Y)$ , conditional on certain observable variables  $Y$ , predicted by the model. The setting is sufficiently general to deal with a wide range of situations. For instance,  $X$  and  $Y$  can be the (latent) volatility and (observable) return processes as is the case of an SV model for asset returns. Or  $X$  and  $Y$  can be the latent state vector and observable yields in a dynamic term structure model.

The posterior distribution  $f(\Psi, X|Y)$  is the main tool to draw inference not only on the coefficient  $\Psi$  but also on the latent vector  $X$ . Since  $f(\Psi, X|Y)$  is unknown in closed-form in relevant applications, MCMC relies on a simulation (a Markov Chain) from the conditional density  $f(\Psi, X|Y)$  to compute mode, mean, and standard deviations for the model coefficients and state variables via the Monte Carlo method.

The posterior  $f(\Psi, X|Y)$  is analytically untractable and extremely high-dimensional, so that simulation directly from  $f(\Psi, X|Y)$  is typically infeasible. The MCMC approach hinges on the Clifford–Hammersley theorem, which determines conditions under which the posterior  $f(\Psi, X|Y)$  is uniquely determined by the marginal posterior distributions  $f(\Psi|X, Y)$  and  $f(X|\Psi, Y)$ . In turn, the posteriors  $f(\Psi|X, Y)$  and  $f(X|\Psi, Y)$  are determined by a set of univariate posterior distributions. Specifically, denote with  $\Psi(i)$  the  $i$ th element of the coefficient  $\Psi$ ,  $i = 1, \dots, K$ , and with  $\Psi(-i)$  the vector consisting of all elements in  $\Psi$  except for the  $i$ th one. Similarly denote with  $X(t)$  the  $t$ th row of the state vector,  $t = 1, \dots, T$ , and with  $X(-t)$  the rest of the vector. Then the Clifford–Hammersley theorem allows to characterize the posterior

$f(\Psi, X|Y)$  via  $K + T$  univariate posteriors,

$$f(\Psi(i)|\Psi(-i), X, Y), \quad i = 1, \dots, K \quad (94)$$

$$f(X(t)|X(-t), \Psi, Y), \quad t = 1, \dots, T. \quad (95)$$

The construction of the Markov Chain relies on the so-called Gibbs sampler. The first step of the algorithm consists in choosing initial values for the coefficient and the state,  $\Psi_0$  and  $X_0$ . When (one of or both) the multi-dimensional posteriors are tractable, the Gibbs sampler generates values  $\Psi_1$  and  $X_1$  directly from  $f(\Psi|X, Y)$  and  $f(X|\Psi, Y)$ . Alternatively, each element of  $\Psi_1$  and  $X_1$  is drawn from the univariate posteriors (94)–(95). Some of these posteriors may also be analytically intractable or efficient algorithms to draw from these posteriors may not exist. In such cases the Metropolis–Hastings algorithm ensures that the simulated sample is consistent with the posterior target distribution. Metropolis–Hastings sampling consists of an accept-reject procedure of the draws from a ‘proposal’ or ‘candidate’ tractable density, which is used to approximate the unknown posterior (see, e.g., Johannes and Polson [187]).

Subsequent iterations of Gibbs sampling, possibly in combination with the Metropolis–Hastings sampling, yield a series of ‘sweeps’  $\{\Psi_s, X_s\}$ ,  $s = 1, \dots, S$ , with limiting distribution  $f(\Psi, X|Y)$ . A long number of sweeps may be necessary to ‘span’ the whole posterior distribution and obtain convergence due to the serial dependence of subsequent draws of coefficients and state variables. When the algorithm has converged, additional simulations provide a sample from the joint posterior distribution.

The MCMC approach has several advantages. First, the inference automatically accounts for parameter uncertainty. Further, the Markov Chain provides a direct and elegant solution to the *smoothing* problem, i.e., the problem of determining the posterior distribution for the state vector  $X$  conditional on the entire data sample,  $f(X(t)|Y(1), \dots, Y(T), \Psi)$ ,  $t = 1, \dots, T$ . The limitation on the approach is largely that efficient sampling schemes for the posterior distribution must be constructed for each specific problem at hand which by nature is case specific and potentially cumbersome or inefficient. Nonetheless, following the development of more general simulation algorithms, the method has proven flexible for efficient estimation of a broad class of important models.

One drawback is that MCMC does not deliver an immediate solution to the *filtering* problem, i.e., determining  $f(X(t)|Y(1), \dots, Y(t), \Psi)$ , and the *forecasting* problem, i.e., determining  $f(X(t+j)|Y(1), \dots, Y(t), \Psi)$ ,  $j > 0$ . However, recent research is overcoming this limitation

through the use of the ‘particle filter’. Bayes rule implies

$$f(X(t+1)|Y(1), \dots, Y(t+1), \Psi) \propto f(Y(t+1)|X(t+1), \Psi)f(X(t+1)|Y(1), \dots, Y(t), \Psi), \quad (96)$$

where the symbol  $\propto$  denotes ‘proportional to’. The first density on the right-hand side of Eq. (96) is determined by the SV model and it is often known in closed form. In contrast, the second density at the far-right end of the equation is given by an integral that involves the unknown filtering density at the prior period,  $f(X(t)|Y(1), \dots, Y(t), \Psi)$ :

$$f(X(t+1)|Y(1), \dots, Y(t), \Psi) = \int f(X(t+1)|X(t), \Psi) \times f(X(t)|Y(1), \dots, Y(t), \Psi) dX(t). \quad (97)$$

The particle method relies on simulations to construct a finite set of weights  $w^i(t)$  and particles  $X^i(t)$ ,  $i = 1, \dots, N$ , that approximate the unknown density with a finite sum,

$$f(X(t)|Y(1), \dots, Y(t), \Psi) \approx \sum_{i=1}^N w^i(t) \delta_{X^i(t)}, \quad (98)$$

where the Dirac function  $\delta_{X^i(t)}$  assigns mass one to the particle  $X^i(t)$ . Once the set of weights and particles are determined, it is possible to re-sample from the discretized distribution. This step yields a simulated sample  $\{X^s(t)\}_{s=1}^S$  which can be used to evaluate the density in Eq. (97) via Monte Carlo integration:

$$f(X(t+1)|Y(1), \dots, Y(t), \Psi) \approx \frac{1}{S} \sum_{s=1}^S f(X(t+1)|X^s(t), \Psi). \quad (99)$$

Equation (99) solves the forecasting problem while combining formulas (96) and (99) solves the filtering problem. The challenge in practical application of the particle filter is to identify an accurate and efficient algorithm to construct the set of particles and weights. We point the interested reader to Kim et al. [193], Pitt and Shephard [228] and Johannes and Polson [187] for a discussion on how to approach this problem.

The usefulness of the MCMC method to solve the inference problem for SV models has been evident since the early work by Jacquier et al. [180], who develop an MCMC algorithm for the logarithmic SV model. Jacquier et al. [181] provide extensions to correlated and non-normal error distributions. Kim et al. [193], Pitt and Shephard [228] and Chib et al. [96] develop simulation-based

methods to solve the filtering problem, while Chib et al. [97] use the MCMC approach to estimate a multivariate SV model. Elerian et al. [135] and Eraker [140] discuss how to extend the MCMC inference method to a continuous-time setting. Eraker [140] uses the MCMC approach to estimate an SV diffusion process for interest rates, while Jones [189] estimates a continuous-time model for the spot rate with non-linear drift function. Eraker et al. [142] estimate an SV jump-diffusion process using data on S&P 500 return while Eraker [141] estimates a similar model using joint data on options and underlying S&P 500 returns. Li et al. [196] allow for Lévy-type jumps in their model. Collin-Dufresne et al. [104] use the MCMC approach to estimate multi-factor affine dynamic term structure model using swap rates data. Johannes and Polson [186] give a comprehensive survey of the still ongoing research on the use of the MCMC approach in the general nonlinear jump-diffusion SV setting.

### Estimation from Option Data

Options’ payoffs are non-linear functions of the underlying security price. This feature renders options highly sensitive to jumps in the underlying price and to return volatility, which makes option data particularly useful to identify return dynamics. As such, several studies have taken advantage of the information contained in option prices, possibly in combination with underlying return data, to estimate SV models with or without discontinuities in returns and volatility.

Applications to derivatives data typically require a model for the pricing errors. A common approach is to posit that the market price of an option,  $O^*$ , normalized by the underlying observed security price  $S^*$ , is the sum of the normalized model-implied option price,  $O/S^*$ , and a disturbance term  $\varepsilon$  (e.g., Renault [230]):

$$\frac{O^*}{S^*} = \frac{O(S^*, V, K, \tau, \Psi)}{S^*} + \varepsilon, \quad (100)$$

where  $V$  is the latent volatility state,  $K$  is the option strike price,  $\tau$  is time to maturity, and  $\Psi$  is the vector with the model coefficients. A pricing error  $\varepsilon$  could arise for several reasons, including measurement error (e.g., price discreteness), asynchronicity between the derivatives and underlying price observations, microstructure effects, and perhaps most importantly specification error. The structure imposed on  $\varepsilon$  depends on the choice of a specific ‘loss function’ used for estimation (e.g., Christoffersen and Jacobs [98]). Several studies have estimated the coefficient vector  $\Psi$  by minimizing the sum of the squared option pricing errors normalized by the underlying price  $S^*$ , as



in Eq. (100). Others have focused on either squared dollar pricing errors, or squared errors normalized by the options market price (instead of  $S^*$ ). The latter approach has the advantage that a \$1 error on an expensive in-the-money option carries less weight than the same error on a cheaper out-of-the-money contract. The drawback is that giving a lot of weight to the pricing errors on short-maturity deep-out-of-the-money options could bias the estimation results. Finally, the common practice of expressing option prices in terms of their Black-Scholes implied volatilities has inspired other scholars to minimize the deviations between Black-Scholes implied volatilities inferred from model and market prices (e.g., Mizrahi [216]). An alternative course is to form a moment-based loss function and follow a GMM- or SMM-type approach to estimate  $\Psi$ . To this end moment conditions stem from distributional assumptions on the pricing error  $\varepsilon$  (e.g.,  $E[\varepsilon] = 0$ ) or from the scores of a reduced-form model that approximates the data.

In estimating the model, some researchers have opted to use a panel of options consisting of contracts with multiple strikes and maturities across dates in the sample period. This choice brings a wealth of information on the cross-sectional and term-structure properties of the implied volatility smirk into the analysis. Others rely on only one option price observation per time period, which shifts the focus to the time-series dimension of the data. Some studies re-estimate the model on a daily basis rather than seeking a single point estimate for the coefficient  $\Psi$  across the entire sample period. This ad hoc approach produces smaller in-sample pricing errors, which can be useful to practitioners, but at the cost of concealing specification flaws by over-fitting the model, which tends to hurt out-of-sample performance. The different approaches are in part dictated by the intended use of the estimated system as practitioners often are concerned with market making and short-term hedging while academics tend to value stable relations that may form the basis for consistent modeling of the dominant features of the system over time.

Early contributions focus on loss functions based on the sum of squared option pricing errors and rely entirely on option data for estimation. This approach typically yields an estimate of the model coefficient  $\Psi$  that embeds an adjustment for risk, i.e., return and volatility dynamics are identified under the risk-neutral rather than the physical probability measure. For instance, Bates [56] considers an SV jump-diffusion model for Deutsche Mark foreign currency options and estimates its coefficient vector  $\Psi$  via nonlinear generalized least squares of the normalized pricing errors with daily option data from Jan-

uary 1984 to June 1991. A similar approach is followed by Bates [57] who fits an SV model with two latent volatility factors and jumps using daily data on options on the S&P 500 futures from January 1988 to December 1993. Bakshi et al. [34] focus on the pricing and hedging of daily S&P 500 index options from June 1988 to May 1991. In their application they re-calibrate the model on a daily basis by minimizing the sum of the squared dollar pricing errors across options with different maturities and strikes. Huang and Wu [173] explore the pricing implications of the time-changed Lévy process by Carr and Wu [84] for daily S&P 500 index options from April 1999 to May 2000. Their Lévy return process allows for discontinuities that exhibit higher jump frequencies compared to the finite-intensity Poisson jump processes in Equations (37)–(41). Further, their model allows for a random time change, i.e., a monotonic transformation of the time variable which generates SV in the diffusion and jump components of returns. In contrast, Bakshi et al. [35] fit an SV jump-diffusion model by SMM using daily data on long-maturity S&P 500 options (LEAPS).

More recent studies have relied on joint data on S&P 500 option prices and underlying index returns, spanning different periods, to estimate the model. This approach forces the same model to price securities in two different markets and relies on information from the derivatives and underlying securities to better pin down model coefficients and risk premia. For instance, Eraker [141] and Jones [190] fit different flavors of the SV model (with and without jumps, respectively) by MCMC. Pan [220] follows a GMM approach to estimate an SV jump-diffusion model using weekly data. She relies on a single at-the-money option price observation each week, which identifies the level of the latent volatility state variable (i.e., at each date she fixes the error term  $\varepsilon$  at zero and solves Eq. (100) for  $V$ ). Aït-Sahalia and Kimmel [7] apply Aït-Sahalia's [4] method to approximate the likelihood function for a joint sample of options and underlying prices. Chernov and Ghysels [93] and Benzoni [59] obtain moment conditions from the scores of a SNP auxiliary model. Similarly, other recent studies have found it useful to use joint derivatives and interest rate data to fit dynamic term structure models, e.g., Almeida et al. [9], and Bikbov and Chernov [62].

Finally, a different literature has studied the option pricing implications of a model in which asset return volatility is a deterministic function of the asset price and time, e.g., Derman and Kani [119], Dupire [134], Rubinstein [233], and Jackwerth and Rubinstein [177]. Since volatility is not stochastic in this setting, we do not review these models here and point the interested reader to, e.g., [133] for an empirical analysis of their performance.

## Future Directions

In spite of much progress in our understanding of volatility new challenges lie ahead. In recent years a wide array of volatility-sensitive products has been introduced. The market for these derivatives has rapidly grown in size and complexity. Research faces the challenge to price and hedge these new products. Moreover, the recent developments in model-free volatility modeling have effectively given empirical content to the latent volatility variable, which opens the way for a new class of estimation methods and specification tests for SV systems. Related, improved volatility measures enable us to shed new light on the properties and implications of the volatility risk premium. Finally, more work is needed to better understand the linkage between fluctuations in economic fundamentals and low- and high-frequency volatility movements. We conclude this chapter by briefly reviewing some open issues in these four areas of research.

## Volatility and Financial Markets Innovation

Volatility is a fundamental input to any financial and real investment decision. Markets have responded to investors' needs by offering an array of volatility-linked instruments. In 1993 the Chicago Board Option Exchange (CBOE) has introduced the VIX index, which measures the market expectations of near-term volatility conveyed by equity-index options. The index was originally computed using the Black-Scholes implied volatilities of eight different S&P 100 option (OEX) series so that, at any given time, it represented the implied volatility of a hypothetical at-the-money OEX option with exactly 30 days to expiration (see [257]). On September 22, 2003, the CBOE began disseminating price level information using a revised 'model-free' method for the VIX index. The new VIX is given by the price of a portfolio of S&P 500 index options and incorporates information from the volatility smirk by using a wider range of strike prices rather than just at-the-money series (see [77]). On March 26, 2004, trading in futures on the VIX Index started on the CBOE Futures Exchange (CFE) while on February 24, 2006, options on the VIX began trading on the Chicago Board Options Exchange. These developments have opened the way for investors to trade on option-implied measures of market volatility. The popularity of the VIX prompted the CBOE to introduce similar indices for other markets, e. g., the VXN NASDAQ 100 Volatility Index.

Along the way, a new over-the-counter market for volatility derivatives has also rapidly grown in size and liquidity. Volatility derivatives are contracts whose payments are expressed as functions of realized variance. Popular ex-

amples are variance swaps, which at maturity pay the difference between realized variance and a fixed strike price. According to estimates by BNP Paribas reported by the Risk [176] magazine, the daily trading volume for variance swaps on indices reached \$4–5 million in vega notional (measured in dollars per volatility point) in 2006, which corresponds to payments in excess of \$1 billion per percentage point of volatility on an annual basis (Carr and Lee [82]). Using variance swaps hedge fund managers and proprietary traders can easily place huge bets on market volatility.

Finally, in recent years credit derivatives markets have evolved in complexity and grown in size. Among the most popular credit derivatives are the credit default swaps (CDS), which provide insurance against the risk of default by a particular company. The buyer of a single-name CDS acquires the right to sell bonds issued by the company at face value when a credit event occurs. Multiple-name contracts can be purchased simultaneously through credit indices. For instance, the CDX indices track the credit spreads for different portfolios of North American companies while the iTraxx Europe indices track the spreads for portfolios of European companies. At the end of 2006 the notional amount of outstanding over-the-counter single- and multi-name CDS contracts stood at \$19 and \$10 trillion, respectively, according to the September 2007 Bank for International Settlements Quarterly Review.

These market developments have raised new interesting issues for research to tackle. The VIX computations based on the new model-free definition of implied volatility used by the CBOE requires the use of options with strike prices that cover the entire support of the return distribution. In practice, liquid options satisfying this requirement often do not exist and the CBOE implementation introduces random noise and systematic error into the index (Jiang and Tian [185]). Related, the VIX implementation entails a truncation, i. e., the CBOE discards illiquid option prices with strikes lying in the tails of the return distribution. As such, the notion of the VIX is more directly linked to that of corridor volatility [26]. In sum, robust implementation of a model free measure of implied volatility is still an open area of research. Future developments in this direction will also have important repercussions on the hedging practices for implied-volatility derivatives.

Pricing and hedging of variance derivatives is another active area of research. Variance swaps admit a simple replication strategy via static positions in call and put options on the underlying asset, similar to model-free implied volatility measures (e. g., [77,83]). In contrast, it is still an open area of research to determine the replication

strategy for derivatives whose payoffs are non-linear function of realized variance, e. g., volatility swaps, which pay the square-root of realized variance, or call and put options on realized variance. [82] is an interesting paper in this direction.

Limited liability gives shareholders the option to default on the firm's debt obligation. As such, a debt claim has features similar to a short position in a put option. The pricing of corporate debt is therefore sensitive to the volatility of the firms' assets: higher volatility increases the probability of default and therefore reduces the price of debt and increases credit spreads. The insights and techniques developed in the SV literature could prove useful in credit risk modeling and applications (e. g., [179,248,260]).

### The Use of Realized Volatility for Estimation of SV Models

Another promising line of research aims at extracting the information in RV measures for the estimation of dynamic asset pricing models. Early work along these lines includes Barndorff-Nielsen and Shephard [51], who decompose RV into actual volatility and realized volatility error. They consider a state-space representation for the decomposition and apply the Kalman filter to estimate different flavors of the SV model. Moreover, Bollerslev and Zhou [68] and Garcia et al. [156], build on the insights of Meddahi [210] to estimate SV diffusion models using conditional moments of integrated volatility. More recently, Todorov [252] generalizes the analysis for the presence of jumps.

Related, recent studies have started to use RV measures to test the implications of models previously estimated with lower-frequency data. Since RV gives empirical content to the latent quadratic variation process, this approach allows for a direct test of the model-implied restrictions on the latent volatility factor. Recent work along these lines includes Andersen and Benzoni [12], who use model-free RV measures to show that the volatility spanning condition embedded in some affine term structure models is violated in the US Treasury market. Christoffersen et al. [100] note that the Heston square-root SV model implies that the dynamics for the standard deviation process are conditionally Gaussian. They reject this condition by examining the distribution of the changes in the square-root RV measure for S&P 500 returns.

### Volatility Risk Premium

More work is needed to better understand the link between asset return volatility and model risk premia. Also in this case, RV measures are a fruitful source of information

to shed new light on the issue. Among the recent studies that pursue this venue is Bollerslev et al. [66], who exploit the moments of RV and option-implied volatility to gauge a measure of the volatility risk premium. Todorov [251] explores the variance risk premium dynamics using high-frequency S&P 500 index futures data and data on the VIX index. He finds the variance risk premium to vary significantly over time and to increase during periods of high volatility and immediately after big jumps in underlying returns. Carr and Wu [85] provide a broader analysis of the variance risk premium for five equity indices and 35 individual stocks. They find the premium to be large and negative for the indices while it is much smaller for the individual stocks. Further, they also find the premium to increase (in absolute value) with the level of volatility. Additional work on the volatility risk premium embedded in individual stock options is in Bakshi and Kapadia [36], Driessen et al. [123], and Duarte and Jones [126]. Other studies have examined the linkage between volatility risk premia and equity returns (e. g., [69]) and hedge-fund performance (e. g., [70]). New research is also examining the pricing of aggregate volatility risk in the cross-section of stock returns. For instance, Ang et al. [30] find that average returns are lower on stocks that have high sensitivities to innovations in aggregate volatility and high idiosyncratic volatility (see also the related work by Chen [90] Ang et al. [32], Bandi et al, Guo et al. [42]). This evidence is consistent with the findings of the empirical option pricing literature, which suggests that there is a negative risk premium for volatility risk. Intuitively, periods of high market volatility are associated to worsened investment opportunities and tend to coincide with negative stock market returns (the so-called leverage effect). As such, investors are willing to pay higher prices (i. e., accept lower expected returns) to hold stocks that do well in high-volatility conditions.

### Determinants of Volatility

Finally, an important area of future research concerns the linkage between asset return volatility and economic uncertainty. Recent studies have proposed general equilibrium models that produce low-frequency fluctuations in conditional volatility, e. g., Campbell and Cochrane [80], Bansal and Yaron [47], McQueen and Vorkink [207], and Tauchen [246]. Related, Engle and Rangel [139] and Engle et al. [138] link macroeconomic variables and long-run volatility movements. It is still an open issue, however, to determine the process through which news about economic fundamentals are embedded into prices to generate high-frequency volatility fluctuations. Early research

by Schwert [236] and Shiller [241] has concluded that the amplitude of the fluctuations in aggregate stock volatility is difficult to explain using simple models of stock valuation. Further, Schwert [236] notes that while aggregate leverage is significantly correlated with volatility, it explains a relatively small part of the movements in stock volatility. Moreover, he finds little evidence that macroeconomic volatility (measured by inflation and industrial production volatility) helps predict future asset return volatility. Model-free realized volatility measures are a useful tool to further investigate this issue. Recent work in this direction includes Andersen et al. [22] and Andersen and Bollerslev [17], who explore the linkage between news arrivals and exchange rates volatility, and Andersen and Benzoni [13], who investigate the determinants of bond yields volatility in the US Treasury market. Related, Balduzzi et al. [38] and Fleming and Remolona [146] study the reaction of trading volume, bid-ask spread, and price volatility to macroeconomic surprises in the US Treasury market, while Brandt and Kavajecz [74] and Pasquariello and Vega [222] focus instead on the price discovery process and explore the implications of order flow imbalances (excess buying or selling pressure) on day-to-day variation in yields.

## Acknowledgments

We are grateful to Olena Chyruk, Bruce Mizrach (the Section Editor) and Neil Shephard for helpful comments and suggestions. Of course, all errors remain our sole responsibility. The views expressed herein are those of the authors and not necessarily those of the Federal Reserve Bank of Chicago or the Federal Reserve System. The work of Andersen is supported by a grant from the NSF to the NBER and support from CREATES funded by the Danish National Research Foundation.

## Bibliography

### Primary Literature

- Ahn DH, Dittmar RF, Gallant AR (2002) Quadratic Term Structure Models: Theory and Evidence. *Rev Finance Stud* 15:243–288
- Ahn DH, Dittmar RF, Gallant AR, Gao B (2003) Purebred or hybrid?: Reproducing the volatility in term structure dynamics. *J Econometrics* 116:147–180
- Ait-Sahalia Y (2002) Maximum-Likelihood Estimation of Discretely-Sampled Diffusions: A Closed-Form Approximation Approach. *Econometrica* 70:223–262
- Ait-Sahalia Y (2007) Closed-Form Likelihood Expansions for Multivariate Diffusions. *Annals of Statistics*, forthcoming
- Ait-Sahalia Y, Hansen LP, Scheinkman J (2004) Operator Methods for Continuous-Time Markov Processes. In: Hansen LP, Ait-Sahalia Y (eds) *Handbook of Financial Econometrics*. North-Holland, Amsterdam, forthcoming
- Ait-Sahalia Y, Jacod J (2007) Testing for Jumps in a Discretely Observed Process. *Annals of Statistics*, forthcoming
- Ait-Sahalia Y, Kimmel R (2007) Maximum Likelihood Estimation of Stochastic Volatility Models. *J Finance Econ* 83:413–452
- Alizadeh S, Brandt MW, Diebold FX (2002) Range-Based Estimation of Stochastic Volatility Models. *J Finance* 57:1047–1091
- Almeida CIR, Graveline JJ, Joslin S (2006) Do Options Contain Information About Excess Bond Returns? Working Paper, UMN, Fundação Getulio Vargas, MIT, Cambridge
- Andersen TG (1994) Stochastic Autoregressive Volatility: A Framework for Volatility Modeling. *Math Finance* 4:75–102
- Andersen TG (1996) Return Volatility and Trading Volume: An Information Flow Interpretation of Stochastic Volatility. *J Finance* 51:169–204
- Andersen TG, Benzoni L (2006) Do Bonds Span Volatility Risk in the US Treasury Market? A Specification Test for Affine Term Structure Models. Working Paper, KSM and Chicago FED, Chicago
- Andersen TG, Benzoni L (2007) The Determinants of Volatility in the US Treasury market. Working Paper, KSM and Chicago FED, Chicago
- Andersen TG, Benzoni L (2007) Realized volatility. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) *Handbook of Financial Time Series*. Springer, Berlin (forthcoming)
- Andersen TG, Benzoni L, Lund J (2002) An Empirical Investigation of Continuous-Time Equity Return Models. *J Finance* 57:1239–1284
- Andersen TG, Benzoni L, Lund J (2004) Stochastic Volatility, Mean Drift and Jumps in the Short Term Interest Rate. Working Paper, Northwestern University, University of Minnesota, and Nykredit Bank, Copenhagen
- Andersen TG, Bollerslev T (1998) Deutsche Mark-Dollar Volatility: Intraday Activity Patterns, Macroeconomic Announcements, and Longer Run Dependencies. *J Finance* 53:219–265
- Andersen TG, Bollerslev T (1998) Answering the Skeptics: Yes, Standard Volatility Models Do Provide Accurate Forecasts. *Int Econ Rev* 39:885–905
- Andersen TG, Bollerslev T, Diebold FX (2004) Parametric and nonparametric volatility measurement. In: Hansen LP, Ait-Sahalia Y (eds) *Handbook of Financial Econometrics*. North-Holland, Amsterdam, forthcoming
- Andersen TG, Bollerslev T, Diebold FX (2007) Roughing It Up: Including Jump Components in Measuring, Modeling and Forecasting Asset Return Volatility. *Rev Econ Statist* 89:701–720
- Andersen TG, Bollerslev T, Diebold FX, Labys P (2003) Modeling and Forecasting Realized Volatility. *Econometrica* 71:579–625
- Andersen TG, Bollerslev T, Diebold FX, Vega C (2003) Micro Effects of Macro Announcements: Real-Time Price Discovery in Foreign Exchange. *Ammer Econom Rev* 93:38–62
- Andersen TG, Bollerslev T, Dobrev D (2007) No-arbitrage semi-martingale restrictions for continuous-time volatility models subject to leverage effects, jumps and i.i.d. noise: Theory and testable distributional implications. *J Econome* 138:125–180



24. Andersen TG, Bollerslev T, Meddahi N (2004) Analytic Evaluation of Volatility Forecasts. *Int Econ Rev* 45:1079–1110
25. Andersen TG, Bollerslev T, Meddahi N (2005) Correcting the Errors: Volatility Forecast Evaluation Using High-Frequency Data and Realized Volatilities. *Econometrica* 73:279–296
26. Andersen TG, Bondarenko O (2007) Construction and Interpretation of Model-Free Implied Volatility. Working Paper, KSM and UIC, Chicago
27. Andersen TG, Chung HJ, Sørensen BE (1999) Efficient Method of Moments Estimation of a Stochastic Volatility Model: A Monte Carlo Study. *J Econom* 91:61–87
28. Andersen TG, Lund J (1997) Estimating continuous-time stochastic volatility models of the short term interest rate diffusion. *J Econom* 77:343–377
29. Ané T, Geman H (2000) Order Flow, Transaction Clock, and Normality of Asset Returns. *J Finance* 55:2259–2284
30. Ang A, Hodrick RJ, Xing Y, Zhang X (2006) The Cross-Section of Volatility and Expected Returns. *J Finance* 51:259–299
31. Ang A, Hodrick RJ, Xing Y, Zhang X (2008) High idiosyncratic volatility and low returns: international and further U.S. evidence. *Finance Econ* (forthcoming)
32. Bachelier L (1900) *Théorie de la Spéculation*. Annales de École Normale Supérieure 3, Gauthier-Villars, Paris. English translation: Cootner PH (ed) (1964) *The Random Character of Stock Market Prices*. MIT Press, Cambridge
33. Back K (1991) Asset Prices for General Processes. *J Math Econ* 20:371–395
34. Bakshi G, Cao C, Chen Z (1997) Empirical Performance of Alternative Option Pricing Models. *J Finance* 52:2003–2049
35. Bakshi G, Cao C, Chen Z (2002) Pricing and hedging long-term options. *J Econom* 94:277–318
36. Bakshi G, Kapadia N (2003) Delta-Hedged Gains and the Negative Market Volatility Risk Premium. *Rev Finance Stud* 16:527–566
37. Bakshi G, Kapadia N, Madan D (2003) Stock Return Characteristics, Skew Laws, and the Differential Pricing of Individual Equity Options. *Rev Finance Stud* 16:101–143
38. Balduzzi P, Elton EJ, Green TC (2001) Economic News and Bond Prices: Evidence from the US Treasury Market. *J Finance Quant Anal* 36:523–543
39. Bandi FM (2002) Short-term interest rate dynamics: a spatial approach. *J Financ Econ* 65:73–110
40. Bandi FM, Moise CE, Russel JR (2008) Market volatility, market frictions, and the cross section of stock returns. Working Paper, University of Chicago, and Case Western Reserve University, Cleveland
41. Bandi FM, Nguyen T (2003) On the functional estimation of jump-diffusion models. *J Econom* 116:293–328
42. Bandi FM, Phillips PCB (2002) Nonstationary Continuous-Time Processes. In: Hansen LP, Ait-Sahalia Y (eds) *Handbook of Financial Econometrics*. North-Holland, Amsterdam, forthcoming
43. Bandi FM, Phillips PCB (2003) Fully nonparametric estimation of scalar diffusion models. *Econometrica* 71:241–283
44. Bandi FM, Phillips PCB (2007) A simple approach to the parametric estimation of potentially nonstationary diffusions. *J Econom* 137:354–395
45. Bandi FM, Russell J (2006) Separating Microstructure Noise from Volatility. *J Finance Econ* 79:655–692
46. Bandi FM, Russell J (2007) Volatility. In: Birge J, Linetsky V (eds) *Handbook of Financial Engineering*. Elsevier, Amsterdam
47. Bansal R, Yaron A (2004) Risks for the Long Run: A Potential Resolution of Asset Pricing Puzzles. *J Finance* 59:1481–1509
48. Bartlett MS (1938) The Characteristic Function of a Conditional Statistic. *J Lond Math Soc* 13:63–67
49. Barndorff-Nielsen OE, Hansen P, Lunde A, Shephard N (2007) Designing Realized Kernels to Measure the Ex-Post Variation of Equity Prices in the Presence of Noise. Working Paper, University of Aarhus, Aarhus. Stanford University, Nuffield College, Oxford
50. Barndorff-Nielsen OE, Shephard N (2001) Non-Gaussian Ornstein-Uhlenbeckbased models and some of their uses in financial economics. *J R Stat Soc B* 63:167–241
51. Barndorff-Nielsen OE, Shephard N (2002) Econometric Analysis of Realised Volatility and its Use in Estimating Stochastic Volatility Models. *J R Stat Soc B* 64:253–280
52. Barndorff-Nielsen OE, Shephard N (2002b) Estimating quadratic variation using realized variance. *J Appl Econom* 17:457–477
53. Barndorff-Nielsen OE, Shephard N (2004) Power and bipower variation with stochastic volatility and jumps. *J Finance Econom* 2:1–37
54. Barndorff-Nielsen OE, Shephard N (2006) Econometrics of testing for jumps in financial economics using bipower variation. *J Finance Econom* 4:1–30
55. Bates DS (1991) The Crash of '87: Was It Expected? The Evidence from Options Markets. *J Finance* 46:1009–1044
56. Bates DS (1996) Jumps and stochastic volatility: exchange rate processes implicit in deutsche mark options. *Rev Finance Stud* 9:69–107
57. Bates DS (2000) Post-'87 crash fears in the S&P 500 futures option market. *J Econom* 94:181–238
58. Bates DS (2006) Maximum Likelihood Estimation of Latent Affine Processes. *Rev Finance Stud* 19:909–965
59. Benzoni L (2002) Pricing Options under Stochastic Volatility: An Empirical Investigation. Working Paper, Chicago FED
60. Benzoni L, Collin-Dufresne P, Goldstein RS (2007) Explaining Pre- and Post-1987 Crash Prices of Equity and Options within a Unified General Equilibrium Framework. Working Paper, Chicago FED, UCB, and UMN, Minneapolis
61. Bibby BM, Jacobsen M, Sørensen M (2004) Estimating Functions for Discretely Sampled Diffusion-Type Models. In: Hansen LP, Ait-Sahalia Y (eds) *Handbook of Financial Econometrics*. North-Holland, Amsterdam, forthcoming
62. Bikbov R, Chernov M (2005) Term Structure and Volatility: Lessons from the Eurodollar Markets. Working Paper, LBS, Deutsche Bank, New York
63. Black F, Scholes M (1973) The Pricing of Options and Corporate Liabilities. *J Political Econ* 81:637–654
64. Bladt M, Sørensen M (2007) Simple Simulation of Diffusion Bridges with Application to Likelihood Inference for Diffusions. Working Paper, University of Copenhagen, Copenhagen
65. Bollen NPB, Whaley RE (2004) Does Net Buying Pressure Affect the Shape of Implied Volatility Functions? *J Finance* 59:711–753
66. Bollerslev T, Gibson M, Zhou H (2004) Dynamic Estimation of Volatility Risk Premia and Investor Risk Aversion from Option-Implied and Realized Volatilities. Working Paper, Duke University, Federal Reserve Board, Washington D.C.



67. Bollerslev T, Jubinsky PD (1999) Equity Trading vol and Volatility: Latent Information Arrivals and Common Long-Run Dependencies. *J Bus Econ Stat* 17:9–21
68. Bollerslev T, Zhou H (2002) Estimating stochastic volatility diffusion using conditional moments of integrated volatility. *J Econom* 109:33–65
69. Bollerslev T, Zhou H (2007) Expected Stock Returns and Variance Risk Premia. Working Paper, Duke University and Federal Reserve Board, Washington D.C.
70. Bondarenko O (2004) Market price of variance risk and performance of hedge funds. Working Paper, UIC, Chicago
71. Brandt MW, Chapman DA (2003) Comparing Multifactor Models of the Term Structure. Working Paper, Duke University and Boston College, Chestnut Hill
72. Brandt MW, Diebold FX (2006) A No-Arbitrage Approach to Range-Based Estimation of Return Covariances and Correlations. *J Bus* 79:61–73
73. Brandt MW, Jones CS (2006) Volatility Forecasting with Range-Based EGARCH Models. *J Bus Econ Stat* 24:470–486
74. Brandt MW, Kavajecz KA (2004) Price Discovery in the US Treasury Market: The Impact of Orderflow and Liquidity on the Yield Curve. *J Finance* 59:2623–2654
75. Brandt MW, Santa-Clara P (2002) Simulated likelihood estimation of diffusions with an application to exchange rate dynamics in incomplete markets. *J Finance Econ* 63:161–210
76. Breidt FJ, Crato N, de Lima P (1998) The detection and estimation of long memory in stochastic volatility. *J Econom* 83:325–348
77. Britten-Jones M, Neuberger A (2000) Option Prices, Implied Price Processes, and Stochastic Volatility. *J Finance* 55:839–866
78. Broadie M, Chernov M, Johannes MJ (2007) Model Specification and Risk Premia: Evidence from Futures Options. *J Finance* 62:1453–1490
79. Buraschi A, Jackwerth J (2001) The price of a smile: hedging and spanning in option markets. *Rev Finance Stud* 14:495–527
80. Campbell JY, Cochrane JH (1999) By Force of Habit: A Consumption-Based Explanation of Aggregate Stock Market Behavior. *J Polit Economy* 107:205–251
81. Campbell JY, Viceira LM (2001) Who Should Buy Long-Term Bonds? *Ammer Econ Rev* 91:99–127
82. Carr P, Lee R (2007) Robust Replication of Volatility Derivatives. Working Paper, NYU and University of Chicago, Chicago
83. Carr P, Madan D (1998) Towards a theory of volatility trading. In: Jarrow R (ed) *Volatility*. Risk Publications, London
84. Carr P, Wu L (2004) Time-changed Lévy processes and option pricing. *J Finance Econ* 71:113–141
85. Carr P, Wu L (2007) Variance Risk Premia. *Rev Finance Stud*, forthcoming
86. Carrasco M, Chernov M, Florens JP, Ghysels E (2007) Efficient estimation of general dynamic models with a continuum of moment conditions. *J Econom* 140:529–573
87. Carrasco M, Florens JP (2000) Generalization of GMM to a continuum of moment conditions. *Econom Theory* 16:797–834
88. Chacko G, Viceira LM (2003) Spectral GMM estimation of continuous-time processes. *J Econom* 116:259–292
89. Chan KC, Karolyi GA, Longstaff FA, Sanders AB (1992) An Empirical Comparison of Alternative Models of the Short-Term Interest Rate. *J Finance* 47:1209–1227
90. Chen J (2003) Intertemporal CAPM and the Cross-Section of Stock Return. Working Paper, USC, Los Angeles
91. Chen RR, Scott L (1993) Maximum likelihood estimation for a multifactor equilibrium model of the term structure of interest rates. *J Fixed Income* 3:14–31
92. Cheridito P, Filipović D, Kimmel RL (2007) Market Price of Risk Specifications for Affine Models: Theory and Evidence. *J Finance Econ*, 83(1):123–170
93. Chernov M, Ghysels E (2002) A study towards a unified approach to the joint estimation of objective and risk neutral measures for the purpose of options valuation. *J Finance Econ* 56:407–458
94. Chernov M, Gallant AR, Ghysels E, Tauchen G (1999) A New Class of Stochastic Volatility Models with Jumps: Theory and Estimation. Working Paper, London Business School, Duke University, University of Northern Carolina
95. Chernov M, Gallant AR, Ghysels E, Tauchen G (2003) Alternative models for stock price dynamics. *J Econom* 116:225–257
96. Chib S, Nardari F, Shephard N (2002) Markov chain Monte Carlo methods for stochastic volatility models. *J Econom* 108:281–316
97. Chib S, Nardari F, Shephard N (2006) Analysis of high dimensional multivariate stochastic volatility models. *J Econom* 134:341–371
98. Christoffersen PF, Jacobs K (2004) The importance of the loss function in option valuation. *J Finance Econ* 72:291–318
99. Christoffersen PF, Jacobs K, Karoui L, Mimouni K (2007) Estimating Term Structure Models Using Swap Rates. Working Paper, McGill University, Montreal
100. Christoffersen PF, Jacobs K, Mimouni K (2006a) Models for S&P 500 Dynamics: Evidence from Realized Volatility, Daily Returns, and Option Prices. Working Paper, McGill University, Montreal
101. Christoffersen PF, Jacobs K, Wang Y (2006b) Option Valuation with Long-run and Short-run Volatility Components. Working Paper, McGill University, Montreal
102. Clark PK (1973) A Subordinated Stochastic Process Model with Finite Variance for Speculative Prices. *Econometrica* 41:135–156
103. Collin-Dufresne P, Goldstein RS (2002) Do Bonds Span the Fixed Income Markets? Theory and Evidence for Unspanned Stochastic Volatility. *J Finance* 57:1685–1730
104. Collin-Dufresne P, Goldstein R, Jones CS (2008) Identification of Maximal Affine Term Structure Models. *J Finance* 63(2):743–795
105. Collin-Dufresne P, Goldstein R, Jones CS (2007b) Can Interest Rate Volatility be Extracted from the Cross Section of Bond Yields? An Investigation of Unspanned Stochastic Volatility. Working Paper, UCB, USC, UMN, Minneapolis
106. Collin-Dufresne P, Solnik B (2001) On the Term Structure of Default Premia in the Swap and LIBOR Markets. *J Finance* 56:1095–1115
107. Comte F, Coutin L, Renault E (2003) Affine Fractional Stochastic Volatility Models with Application to Option Pricing. Working Paper, CIRANO, Montreal
108. Comte F, Renault E (1998) Long memory in continuous-time stochastic volatility models. *Math Finance* 8:291–323
109. Conley TG, Hansen LP, Luttmer EGJ, Scheinkman JA (1997) Short-term interest rates as subordinated diffusions. *Rev Finance Stud* 10:525–577

110. Corsi F (2003) A Simple Long Memory Model of realized Volatility. Working paper, University of Southern Switzerland, Lugano
111. Coval JD, Shumway T (2001) Expected Option Returns. *J Finance* 56:983–1009
112. Cox JC, Ingersoll JE, Ross SA (1985) An Intertemporal General Equilibrium Model of Asset Prices. *Econometrica* 53:363–384
113. Cox JC, Ingersoll JE, Ross SA (1985) A Theory of the Term Structure of Interest Rates. *Econometrica* 53:385–407
114. Dai Q, Singleton KJ (2000) Specification Analysis of Affine Term Structure Models. *J Finance* 55:1943–1978
115. Dai Q, Singleton KJ (2003) Term Structure Dynamics in Theory and Reality. *Rev Finance Stud* 16:631–678
116. Danielsson J (1994) Stochastic Volatility in Asset Prices: Estimation by Simulated Maximum Likelihood. *J Econom* 64:375–400
117. Danielsson J, Richard JF (1993) Accelerated Gaussian Importance Sampler with Application to Dynamic Latent Variable Models. *J Appl Econom* 8:S153–S173
118. Deo R, Hurvich C (2001) On the Log Periodogram Regression Estimator of the Memory Parameter in Long Memory Stochastic Volatility Models. *Econom Theory* 17:686–710
119. Derman E, Kani I (1994) The volatility smile and its implied tree. Quantitative Strategies Research Notes, Goldman Sachs, New York
120. Diebold FX, Nerlove M (1989) The Dynamics of Exchange Rate Volatility: A Multivariate Latent Factor ARCH Model. *J Appl Econom* 4:1–21
121. Diebold FX, Strasser G (2007) On the Correlation Structure of Microstructure Noise in Theory and Practice. Working Paper, University of Pennsylvania, Philadelphia
122. Dobrev D (2007) Capturing Volatility from Large Price Moves: Generalized Range Theory and Applications. Working Paper, Federal Reserve Board, Washington D.C.
123. Driessen J, Maenhout P, Vilkov G (2006) Option-Implied Correlations and the Price of Correlation Risk. Working Paper, University of Amsterdam and INSEAD, Amsterdam
124. Duarte J (2004) Evaluating An Alternative Risk Preference in Affine Term Structure Models. *Rev Finance Stud* 17:370–404
125. Duarte J (2007) The Causal Effect of Mortgage Refinancing on Interest-Rate Volatility: Empirical Evidence and Theoretical Implications. *Rev Finance Stud*, forthcoming
126. Duarte J, Jones CS (2007) The Price of Market Volatility Risk. Working Paper, University of Washington and USC, Washington
127. Duffee GR (2002) Term Premia and Interest Rate Forecasts in Affine Models. *J Finance* 57:405–443
128. Duffee G, Stanton R (2008) Evidence on simulation inference for near unit-root processes with implications for term structure estimation. *J Finance Econom* 6:108–142
129. Duffie D, Kan R (1996) A yield-factor model of interest rates. *Math Finance* 6:379–406
130. Duffie D, Pan J, Singleton KJ (2000) Transform Analysis and Asset Pricing for Affine Jump-Diffusions. *Econometrica* 68:1343–1376
131. Duffie D, Singleton KJ (1993) Simulated Moments Estimation of Markov Models of Asset Prices. *Econometrica* 61:929–952
132. Duffie D, Singleton KJ (1997) An Econometric Model of the Term Structure of Interest-Rate Swap Yields. *J Finance* 52:1287–1321
133. Dumas B, Fleming J, Whaley RE (1996) Implied Volatility Functions: Empirical Tests. *J Finance* 53:2059–2106
134. Dupire B (1994) Pricing with a smile. *Risk* 7:18–20
135. Elerian O, Chib S, Shephard N (2001) Likelihood Inference for Discretely Observed Nonlinear Diffusions. *Econometrica* 69:959–994
136. Engle RF (2002) New frontiers for ARCH models. *J Appl Econom* 17:425–446
137. Engle RF, Gallo GM (2006) A multiple indicators model for volatility using intra-daily data. *J Econom* 131:3–27
138. Engle RF, Ghysels E, Sohn B (2006) On the Economic Sources of Stock Market Volatility. Working Paper, NYU and UNC, Chapel Hill
139. Engle RF, Rangel JG (2006) The Spline-GARCH Model for Low Frequency Volatility and Its Global Macroeconomic Causes. Working Paper, NYU, New York
140. Eraker B (2001) MCMC Analysis of Diffusions with Applications to Finance. *J Bus Econ Stat* 19:177–191
141. Eraker B (2004) Do Stock Prices and Volatility Jump? Reconciling Evidence from Spot and Option Prices. *J Finance* 59:1367–1404
142. Eraker B, Johannes MS, Polson N (2003) The Impact of Jumps in Volatility and Returns. *J Finance* 58:1269–1300
143. Fan R, Gupta A, Ritchken P (2003) Hedging in the Possible Presence of Unspanned Stochastic Volatility: Evidence from Swaption Markets. *J Finance* 58:2219–2248
144. Fiorentina G, Sentana E, Shephard N (2004) Likelihood-Based Estimation of Latent Generalized ARCH Structures. *Econometrica* 72:1481–1517
145. Fisher M, Gilles C (1996) Estimating exponential-affine models of the term structure. Working Paper, Atlanta FED, Atlanta
146. Fleming MJ, Remolona EM (1999) Price Formation and Liquidity in the US Treasury Market: The Response to Public Information. *J Finance* 54:1901–1915
147. Forsberg L, Ghysels E (2007) Why Do Absolute Returns Predict Volatility So Well? *J Finance Econom* 5:31–67
148. French KR, Schwert GW, Stambaugh RF (1987) Expected stock returns and volatility. *J Finance Econ* 19:3–29
149. Fridman M, Harris L (1998) A Maximum Likelihood Approach for Non-Gaussian Stochastic Volatility Models. *J Bus Econ Stat* 16:284–291
150. Gallant AR, Hsieh DA, Tauchen GE (1997) Estimation of Stochastic Volatility Models with Diagnostics. *J Econom* 81:159–192
151. Gallant AR, Hsu C, Tauchen GE (1999) Using Daily Range Data to Calibrate Volatility Diffusions and Extract the Forward Integrated Variance. *Rev Econ Stat* 81:617–631
152. Gallant AR, Long JR (1997) Estimating stochastic differential equations efficiently by minimum chi-squared. *Biometrika* 84:125–141
153. Gallant AR, Rossi PE, Tauchen GE (1992) Stock Prices and Volume. *Rev Finance Stud* 5:199–242
154. Gallant AR, Tauchen GE (1996) Which Moments to Match. *Econ Theory* 12:657–681
155. Gallant AR, Tauchen G (1998) Reprojecting Partially Observed Systems With Application to Interest Rate Diffusions. *J Amer Stat Assoc* 93:10–24
156. Garcia R, Lewis MA, Pastorello S, Renault E (2001) Estimation of Objective and Risk-neutral Distributions based on Moments of Integrated Volatility, Working Paper. Univer-

- sité de Montréal, Banque Nationale du Canada, Università di Bologna, UNC
157. Garman MB, Klass MJ (1980) On the Estimation of Price Volatility From Historical Data. *J Bus* 53:67–78
  158. Ghysels E, Harvey AC, Renault E (1996) Stochastic Volatility. In: Maddala GS, Rao CR (eds) *Handbook of Statistics*, vol 14. North Holland, Amsterdam
  159. Ghysels E, Santa-Clara P, Valkanov R (2006) Predicting Volatility: How to Get the Most Out of Returns Data Sampled at Different Frequencies. *J Econom* 131:59–95
  160. Glosten LR, Jagannathan R, Runkle D (1993) On the Relation Between the Expected Value and the Volatility of the Nominal Excess Return on Stocks. *J Finance* 48:1779–1801
  161. Gong FF, Remolona EM (1996) A three-factor econometric model of the US term structure. Working paper, Federal Reserve Bank of New York, New York
  162. Gouriéroux C, Monfort A, Renault E (1993) Indirect Inference. *J Appl Econom* 8:S85–S118
  163. Guo H, Neely CJ, Higbee J (2007) Foreign Exchange Volatility Is Priced in Equities. *Finance Manag*, forthcoming
  164. Han B (2007) Stochastic Volatilities and Correlations of Bond Yields. *J Finance* 62:1491–1524
  165. Hansen PR, Lunde A (2006) Realized Variance and Market Microstructure Noise. *J Bus Econ Stat* 24:127–161
  166. Harvey AC (1998) Long memory in stochastic volatility. In: Knight J, Satchell S (eds) *Forecasting Volatility in Financial Markets*. Butterworth-Heinemann, London
  167. Harvey AC, Ruiz E, Shephard N (1994) Multivariate Stochastic Variance Models. *Rev Econ Stud* 61:247–264
  168. Harvey AC, Shephard N (1996) Estimation of an Asymmetric Stochastic Volatility Model for Asset Returns. *J Bus Econ Stat* 14:429–434
  169. Heidari M, Wu L (2003) Are Interest Rate Derivatives Spanned by the Term Structure of Interest Rates? *J Fixed Income* 13:75–86
  170. Heston SL (1993) A closed-form solution for options with stochastic volatility with applications to bond and currency options. *Rev Finance Stud* 6:327–343
  171. Ho M, Perraudin W, Sørensen BE (1996) A Continuous Time Arbitrage Pricing Model with Stochastic Volatility and Jumps. *J Bus Econ Stat* 14:31–43
  172. Huang X, Tauchen G (2005) The relative contribution of jumps to total price variation. *J Finance Econom* 3:456–499
  173. Huang J, Wu L (2004) Specification Analysis of Option Pricing Models Based on Time-Changed Levy Processes. *J Finance* 59:1405–1440
  174. Hull J, White A (1987) The Pricing of Options on Assets with Stochastic Volatilities. *J Finance* 42:281–300
  175. Hurvich CM, Soulier P (2007) Stochastic Volatility Models with Long Memory. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) *Handbook of Financial Time Series*. Springer, Berlin
  176. Jung J (2006) Vexed by variance. *Risk August*
  177. Jackwerth JC, Rubinstein M (1996) Recovering Probability Distributions from Option Prices. *J Finance* 51:1611–1631
  178. Jacobs K, Karoui L (2007) Conditional Volatility in Affine Term Structure Models: Evidence from Treasury and Swap Markets. Working Paper, McGill University, Montreal
  179. Jacobs K, Li X (2008) Modeling the Dynamics of Credit Spreads with Stochastic Volatility. *Manag Sci* 54:1176–1188
  180. Jacquier E, Polson NG, Rossi PE (1994) Bayesian Analysis of Stochastic Volatility Models. *J Bus Econ Stat* 12:371–389
  181. Jacquier E, Polson NG, Rossi PE (2004) Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *J Econom* 122:185–212
  182. Jagannathan R, Kaplin A, Sun S (2003) An evaluation of multi-factor CIR models using LIBOR, swap rates, and cap and swap prices. *J Econom* 116:113–146
  183. Jarrow R, Li H, Zhao F (2007) Interest Rate Caps “Smile” Too! But Can the LIBOR Market Models Capture the Smile? *J Finance* 62:345–382
  184. Jiang GJ, Knight JL (2002) Efficient Estimation of the Continuous Time Stochastic Volatility Model via the Empirical Characteristic Function. *J Bus Econ Stat* 20:198–212
  185. Jiang GJ, Tian YS (2005) The Model-Free Implied Volatility and Its Information Content. *Rev Finance Stud* 18:1305–1342
  186. Johannes MS, Polson N (2003) MCMC Methods for Continuous-Time Financial Econometrics. In: Hansen LP, Ait-Sahalia I (eds) *Handbook of Financial Econometrics*. North-Holland, Amsterdam, forthcoming
  187. Johannes MS, Polson N (2006) Particle Filtering. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) *Handbook of Financial Time Series*. Springer, Berlin
  188. Johnson H, Shanno D (1987) Option Pricing when the Variance Is Changing. *J Finance Quant Anal* 22:143–152
  189. Jones CS (2003) Nonlinear Mean Reversion in the Short-Term Interest Rate. *Rev Finance Stud* 16:793–843
  190. Jones CS (2003) The dynamics of stochastic volatility: evidence from underlying and options markets. *J Econom* 116:181–224
  191. Joslin S (2006) Can Unspanned Stochastic Volatility Models Explain the Cross Section of Bond Volatilities? Working Paper, MIT, Cambridge
  192. Joslin S (2007) Pricing and Hedging Volatility Risk in Fixed Income Markets. Working Paper, MIT, Cambridge
  193. Kim SN, Shephard N, Chib S (1998) Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models. *Rev Econ Stud* 65:361–393
  194. Lamoureux CG, Lastrapes WD (1994) Endogenous Trading vol and Momentum in Stock-Return Volatility. *J Bus Econ Stat* 14:253–260
  195. Lee SS, Mykland PA (2006) Jumps in Financial Markets: A New Nonparametric Test and Jump Dynamics. Working Paper, Georgia Institute of Technology and University of Chicago, Chicago
  196. Li H, Wells MT, Yu CL (2006) A Bayesian Analysis of Return Dynamics with Lévy Jumps. *Rev Finance Stud*, forthcoming
  197. Li H, Zhao F (2006) Unspanned Stochastic Volatility: Evidence from Hedging Interest Rate Derivatives. *J Finance* 61:341–378
  198. Liesenfeld R (1998) Dynamic Bivariate Mixture Models: Modeling the Behavior of Prices and Trading Volume. *J Bus Econ Stat* 16:101–109
  199. Liesenfeld R (2001) A Generalized Bivariate Mixture Model for Stock Price Volatility and Trading Volume. *J Econom* 104:141–178
  200. Liesenfeld R, Richard J-F (2003) Univariate and Multivariate Stochastic Volatility Models: Estimation and Diagnostics. *J Empir Finance* 10:505–531
  201. Litterman R, Scheinkman JA (1991) Common Factors Affecting Bond Returns. *J Fixed Income* 1:54–61
  202. Liu C, Maheu JM (2007) Forecasting Realized Volatility: A Bayesian Model Averaging Approach. Working Paper, University of Toronto, Toronto

203. Lo AW (1988) Maximum likelihood estimation of generalized Itô processes with discretely-sampled data. *Econ Theory* 4:231–247
204. Longstaff FA, Schwartz ES (1992) Interest Rate Volatility and the Term Structure: A Two-Factor General Equilibrium Model. *J Finance* 47:1259–1282
205. McAleer M, Medeiros MC (2007) Realized Volatility: A Review. *Econ Rev*, forthcoming
206. McFadden D (1989) A Method of Simulated Moments for Estimation of Discrete Response Models Without Numerical Integration. *Econometrica* 57:995–1026
207. McQueen G, Vorkink K (2004) Whence GARCH? A Preference-Based Explanation for Conditional Volatility. *Rev Finance Stud* 17:915–949
208. Meddahi N (2001) An Eigenfunction Approach for Volatility Modeling. Working Paper, Imperial College, London
209. Meddahi N (2002) A Theoretical Comparison Between Integrated and Realized Volatility. *J Appl Econom* 17:475–508
210. Meddahi N (2002) Moments of Continuous Time Stochastic Volatility Models. Working Paper, Imperial College, London
211. Melino A, Turnbull SM (1990) Pricing foreign currency options with stochastic volatility. *J Econom* 45:239–265
212. Merton RC (1969) Lifetime portfolio selection under uncertainty: the continuous-time case. *Rev Econ Stat* 51:247–257
213. Merton RC (1973) An Intertemporal Capital Asset Pricing Model. *Econometrica* 41:867–887
214. Merton RC (1976) Option pricing when underlying stock returns are discontinuous. *J Finance Econ* 3:125–144
215. Merton RC (1980) On estimating the expected return on the market: An exploratory investigation. *J Finance Econ* 8:323–361
216. Mizrahi B (2006) The Enron Bankruptcy: When Did The Options Market Lose Its Smirk. *Rev Quant Finance Acc* 27:365–382
217. Mizrahi B (2007) Recovering Probabilistic Information From Options Prices and the Underlying. In: Lee C, Lee AC (eds) *Handbook of Quantitative Finance*. Springer, New York
218. Nelson DB (1991) Conditional Heteroskedasticity in Asset Returns: A New Approach. *Econometrica* 59:347–370
219. Nelson DB, Foster DP (1994) Asymptotic Filtering Theory for Univariate ARCH Models. *Econometrica* 62:1–41
220. Pan J (2002) The jump-risk premia implicit in options: evidence from an integrated time-series study. *J Finance Econ* 63:3–50
221. Parkinson M (1980) The Extreme Value Method for Estimating the Variance of the Rate of Return. *J Bus* 53:61–65
222. Pasquariello P, Vega C (2007) Informed and Strategic Order Flow in the Bond Markets. *Rev Finance Stud* 20:1975–2019
223. Pearson ND, Sun TS (1994) Exploiting the Conditional Density in Estimating the Term Structure: An Application to the Cox, Ingersoll, and Ross Model. *J Finance* 49:1279–1304
224. Pedersen AR (1995) A new approach to maximum likelihood estimation for stochastic differential equations based on discrete observations. *Scand J Stat* 22:55–71
225. Pennacchi GG (1991) Identifying the Dynamics of Real Interest Rates and Inflation: Evidence Using Survey Data. *Rev Finance Stud* 4:53–86
226. Piazzesi M (2003) Affine Term Structure Models. In: Hansen LP, Ait-Sahalia I (eds) *Handbook of Financial Econometrics*. North-Holland, Amsterdam, forthcoming
227. Piazzesi M (2005) Bond Yields and the Federal Reserve. *J Political Econ* 113:311–344
228. Pitt MK, Shephard N (1999) Filtering via simulation: auxiliary particle filter. *J Am Stat Assoc* 94:590–599
229. Protter P (1992) *Stochastic Integration and Differential Equations: A New Approach*. Springer, New York
230. Renault E (1997) Econometric models of option pricing errors. In: Kreps D, Wallis K (eds) *Advances in Economics and Econometrics, Seventh World Congress*. Cambridge University Press, New York, pp 223–278
231. Richardson M, Smith T (1994) A Direct Test of the Mixture of Distributions Hypothesis: Measuring the Daily Flow of Information. *J Financ Quant Anal* 29:101–116
232. Rosenberg B (1972) The Behavior of Random Variables with Nonstationary Variance and the Distribution of Security Prices. working Paper, UCB, Berkeley
233. Rubinstein M (1994) Implied Binomial Trees. *J Financ* 49:771–818
234. Santa-Clara P (1995) Simulated likelihood estimation of diffusions with an application to the short term interest rate. Dissertation, INSEAD
235. Schaumburg E (2005) Estimation of Markov processes with Levy type generators. Working Paper, KSM, Evanston
236. Schwert GW (1989) Why Does Stock Market Volatility Change Over Time? *J Financ* 44:1115–1153
237. Schwert GW (1990) Stock Volatility and the Crash of '87. *Rev Financ Stud* 3:77–102
238. Scott LO (1987) Option Pricing when the Variance Changes Randomly: Theory, Estimation and an Application. *J Financ Quant Anal* 22:419–438
239. Shephard N (1996) Statistical Aspects of ARCH and Stochastic Volatility Models. In: Cox DR, Hinkley DV, Barndorff-Nielsen OE (eds) *Time Series Models in Econometrics, Finance and Other Fields*. Chapman & Hall, London, pp 1–67
240. Shephard N (2004) *Stochastic Volatility: Selected Readings*. Oxford University Press, Oxford
241. Shiller RJ (1981) Do Stock Prices Move Too Much to be Justified by Subsequent Changes in Dividends? *Ammer Econ Rev* 71:421–436
242. Singleton KJ (2001) Estimation of affine asset pricing models using the empirical characteristic function. *J Econom* 102:111–141
243. Smith AA Jr (1993) Estimating Nonlinear Time-series Models using Simulated Vector Autoregressions. *J Appl Econom* 8:S63–S84
244. Stein JC (1989) Overreactions in the Options Market. *J Finance* 44:1011–1023
245. Stein EM, Stein JC (1991) Stock price distributions with stochastic volatility: an analytic approach. *Rev Finance Stud* 4:727–752
246. Tauchen GE (2005) Stochastic Volatility in General Equilibrium. Working Paper, Duke University Durham
247. Tauchen GE, Pitts M (1983) The Price Variability-Volume Relationship on Speculative Markets. *Econometrica* 51:485–505
248. Tauchen GE, Zhou H (2007) Realized Jumps on Financial Markets and Predicting Credit Spreads. Working Paper, Duke University and Board of Governors, Washington D.C.
249. Taylor SJ (1986) *Modeling Financial Time Series*. Wiley, Chichester



250. Thompson S (2004) Identifying Term Structure Volatility from the LIBOR-Swap Curve. Working Paper, Harvard University, Boston
251. Todorov V (2006) Variance Risk Premium Dynamics. Working Paper, KSM, Evanston
252. Todorov V (2006) Estimation of Continuous-time Stochastic Volatility Models with Jumps using High-Frequency Data. Working Paper, KSM, Evanston
253. Trolle AB, Schwartz ES (2007) Unspanned stochastic volatility and the pricing of commodity derivatives. Working Paper, Copenhagen Business School and UCLA, Copenhagen
254. Trolle AB, Schwartz ES (2007) A general stochastic volatility model for the pricing of interest rate derivatives. *Rev Finance Stud*, forthcoming
255. Vasicek OA (1977) An equilibrium characterization of the term structure. *J Finance Econ* 5:177–188
256. Wiggins JB (1987) Option Values under Stochastic Volatility: Theory and Empirical Estimates. *J Finance Econ* 19:351–372
257. Whaley RE (1993) Derivatives on Market Volatility: Hedging Tools Long Overdue. *J Deriv* 1:71–84
258. Wright J, Zhou H (2007) Bond Risk Premia and Realized Jump Volatility. Working Paper, Board of Governors, Washington D.C.
259. Yang D, Zhang Q (2000) Drift-Independent Volatility Estimation Based on High, Low, Open, and Close Prices. *J Bus* 73:477–491
260. Zhang BY, Zhou H, Zhu H (2005) Explaining Credit Default Swap Spreads with the Equity Volatility and Jump Risks of Individual Firms. Working Paper, Fitch Ratings, Federal Reserve Board, BIS
261. Zhang L (2007) What you don't know cannot hurt you: On the detection of small jumps. Working Paper, UIC, Chicago
262. Zhang L, Mykland PA, Ait-Sahalia Y (2005) A Tale of Two Time Scales: Determining Integrated Volatility with Noisy High Frequency Data. *J Ammer Stat Assoc* 100:1394–1411
- Jungbacker B, Koopman SJ (2007) Parameter Estimation and Practical Aspect of Modeling Stochastic Volatility. Working Paper, Vrije Universiteit, Amsterdam
- Mykland PA (2003) Option Pricing Bounds and Statistical Uncertainty. In: Hansen LP, Ait-Sahalia Y (eds) *Handbook of Financial Econometrics*. North-Holland, Amsterdam, forthcoming
- Renault E (2007) Moment-Based Estimation of Stochastic Volatility Models. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) *Handbook of Financial Time Series*. Springer, Berlin
- Singleton KJ (2006) *Empirical Dynamic Asset Pricing: Model Specification and Econometric Assessment*. Princeton University Press, Princeton

---

## Structurally Dynamic Cellular Automata

ANDREW ILACHINSKI

Center for Naval Analyses, Alexandria, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[The Basic Model](#)

[Emerging Patterns and Behaviors](#)

[SDCA as Models of Computation](#)

[Generalized SDCA Models](#)

[Related Graph Dynamical Systems](#)

[SDCA as Models of Fundamental Physics](#)

[Future Directions and Speculations](#)

[Bibliography](#)

### Books and Reviews

- Asai M, McAleer M, Yu J (2006) Multivariate Stochastic Volatility: A Review. *Econom Rev* 25:145–175
- Bates DS (2003) Empirical option pricing: a retrospection. *J Econom* 116:387–404
- Campbell JY, Lo AW, MacKinlay AC (1996) *The Econometrics of Financial Markets*. Princeton University Press, Princeton
- Chib S, Omori Y, Asai M (2007) Multivariate Stochastic Volatility. Working Paper, Washington University
- Duffie D (2001) *Dynamic Asset Pricing Theory*. Princeton University Press, Princeton
- Gallant AR, Tauchen G (2002) Simulated Score Methods and Indirect Inference for Continuous-time Models. In: Hansen LP, Ait-Sahalia Y (eds) *Handbook of Financial Econometrics*. North-Holland, Amsterdam, forthcoming
- Garcia R, Ghysels E, Renault E (2003) The Econometrics of Option Pricing. In: Hansen LP, Ait-Sahalia Y (eds) *Handbook of Financial Econometrics*. North-Holland, Amsterdam, forthcoming
- Gouriéroux C, Jasiak J (2001) *Financial Econometrics*. Princeton University Press, Princeton
- Johannes MS, Polson N (2006) Markov Chain Monte Carlo. In: Andersen TG, Davis RA, Kreiss JP, Mikosch T (eds) *Handbook of Financial Time Series*. Springer, Berlin

### Glossary

**Adjacency matrix** The *adjacency matrix* of a graph with  $N$  sites is an  $N \times N$  matrix  $[a_{ij}]$  with entries  $a_{ij} = 1$  if  $i$  and  $j$  are linked, and  $a_{ij} = 0$  otherwise. The adjacency matrix is symmetric ( $a_{ij} = a_{ji}$ ) if the links in the graph are undirected.

**Coupler link rules** *Coupler rules* are local rules that act on pairs of *next-nearest sites* of a graph at time  $t$  to decide whether they should be linked at  $t + 1$ . The decision rules fall into one of three basic classes – totalistic (T), outer-totalistic (OT) or restricted-totalistic (RT) – but can be as varied as those for conventional cellular automata.

**Decoupler link rules** *Decoupler rules* are local rules that act on pairs of *linked sites* of a graph at time  $t$  to decide whether they should be unlinked at  $t + 1$ . As for



coupler rules, the decision rules fall into one of three basic classes – totalistic (T), outer-totalistic (OT) or restricted-totalistic (RT) – but can be as varied as those for conventional cellular automata.

**Degree** The *degree* of a node (or site,  $i$ ) of a graph is equal to the number of distinct nodes to which  $i$  is linked, and where the links are assumed to possess no directional information. In general graphs, the *in-degree* (= number of incoming links towards  $i$ ) is distinguished from the *out-degree* (= number of outgoing links originating at  $i$ ).

**Effective dimension** A quantity used to approximate the *dimensionality* of a graph. It is defined as the ratio between the average number of next-nearest neighbors to the average degree, both averaged over all nodes of the graph. The effective dimension equals the Euclidean dimension  $d$ , in cases where the graph is the familiar  $d$ -dimensional hypercubic lattice.

**Graph** A *graph* is a finite, nonempty set of nodes (referred to as “sites” throughout this article), together with (a possibly empty) set of edges (or links). The links may be either *directed* (in which case the edge from a site  $i$ , say, is directed away from  $i$  toward another site  $j$ , and is considered distinct from another directed edge originating at  $j$  and pointed toward  $i$ ) or *undirected* (in which case if a link exists between sites  $i$  and  $j$  it carries no directional information).

**Graph grammar** *Graph grammars* (sometimes also referred to as *graph rewriting systems*) apply formal language theory to networks. Each language specifies the space of “valid structures”, and the production (or “rewrite”) rules by which given graphs may be transformed into other valid graphs.

**Graph metric function** The *graph metric function* defines the distance between any two nodes,  $i$  and  $j$ . It is equal to the length of the shortest path between  $i$  and  $j$ . If no path exists (such as when  $i$  and  $j$  are on two disconnected components of the same graph), the distance is assumed to be equal to  $\infty$ .

**Graph-rewriting automata** *Graph-rewriting automata* are generalized CA-like systems in which both (the number of) nodes and links are allowed to change.

**Next-nearest neighbor** Two sites  $i$  and  $j$  are *next-nearest neighbors* in a graph if (1) they are not directly linked (so that  $a_{ij} = 0$ ; see *adjacency matrix*), and (2) there exists at least one other site  $k$  such that  $k \notin \{i, j\}$ , and  $i$  and  $j$  are both lined to  $k$ .

**Random dynamics approximation** The long-term behavior of structurally dynamic cellular automata may be approximated in certain cases (in which the structure and value configurations are both sufficiently ran-

dom and uncorrelated) by a *random dynamics approximation*: values of sites are replaced by the probability  $p_\sigma$  of a site having value  $\sigma$  (and is assumed to be equal for all sites), and links between sites are replaced by the probability  $p_\ell$  of being linked (and also assumed to be the same for all pairs of sites). The approximation often yields qualitatively correct predictions about how the real system evolves under a specific set of rules; for example, to predict whether one expects unbounded growth or that the lattice will eventually settle onto a low periodic state or simply decay.

**Restricted totalistic rules** *Restricted totalistic rules* are a generalized class of link rules (operating on pairs of sites,  $i$  and  $j$ ), analogous to “outer totalistic” rules (that operate on site values) used in conventional CA. The local neighborhood around  $i$  and  $j$  is first partitioned into three sets: (1) the two sites,  $i$  and  $j$ ; (2) sites connected to either  $i$  or  $j$ , but not both; and (3) sites connected to both  $i$  and  $j$ . The *restricted totalistic rule* is then completely defined by associating a specific action with each possible 3-tuple of site-value sums (where the individual components represent a unique sum in each of the three neighborhoods).

**Structurally dynamic cellular automata** *Structurally dynamic cellular automata* are generalizations of conventional cellular automata models in which the underlying lattice structure is dynamically coupled to the local site-value configurations.

**SDCA model hierarchy** The *SDCA model hierarchy* is a set of eight related structurally dynamic cellular automata models, defined explicitly for studying their formal computational capabilities. The hierarchy is ordered (from lowest to highest level) according to their relative computational strength. For example, the SDCA model at the top of the hierarchy is capable of simulating a conventional CA with a speedup factor of two.

## Definition of the Subject

*Structurally dynamic cellular automata* (abbreviated, SDCA) are a generalized class of CA in which the topological structure of the (usually quiescent) underlying lattice is dynamically coupled to the local site value configuration. The coupling is defined to treat *geometry* and *value configurations* on an approximately equal footing: the lattice structure is altered locally as a function of individual site neighborhood value-states and geometries, while the underlying local topology supports site-value evolution precisely as in conventional nearest-neighbor CA models defined on random lattices.

SDCA provide a dynamical framework for a CA-like analysis of the generation, transmission and interaction of topological disturbances in a lattice. Moreover, they provide a natural testbed for studying self organized geometry; by which we mean true structural evolution, and not merely space-time patterns of value configurations that may be *interpreted* geometrically (but are really just “bits” of information overlayed on top of an otherwise static background lattice).

## Introduction

SDCA were formally introduced in 1986 as part of a physics doctoral dissertation by Ilachinski [31], and developed further by Ilachinski and Halpern [29,30], Halpern [21,23], Halpern and Caltagirone [22], Majercik [39], and Alonso-Sanz and Martín [6,7,8]; in their original incarnation [28], and at least two subsequent papers [22,61], SDCA were called *topological automata*. Pedagogical discussions appear in Adamatzky [1] and Ilachinski [32]. Extensions of the basic SDCA model (all discussed in this article) include the addition of probabilistic rules, memory and reversibility.

Applications include the simulation of crystal growth [36], the study of pattern formation of random cellular structures [66], modeling synaptic plasticity in neural network models [19], phase transitions in chemical systems [62], chemical self-assembly [26], and gene-regulatory networks [22]. Majercik [39] has studied SDCA as generalized models of computation, and describes a CA-universal SDCA that can simulate any conventional CA of the same dimension.

More recently, O’Sullivan [56] and Saidani [63,64] have used graph-based CA models similar to SDCA to study urban dynamics and emergent behaviors of self-reconfigurable robots, respectively. Tomita et al. [67,68,69,70,71] have introduced *graph-rewriting automata* in which both links and (the number of) nodes are allowed to change; and show that these systems are capable of both self-replication and Turing universality (among with many other emergent behaviors). Since SDCA provide the basic formalism for describing locally induced topological changes within arbitrary graphs, they are a potentially powerful general tool for studying complex adaptive networks, such as communication and social networks [6]. The *concept* behind SDCA has also been used as a foundation for philosophical musings about computationally emergent artificiality [49].

More ambitious applications of SDCA encroach on fundamental physics. Because SDCA are inherently self-modifying systems – in which physical events are not just

dynamically coupled to, but are an integral part of the spatio-temporal arena on which their transformations are defined – they are a potentially powerful methodological and ontological tool for exploring discrete pre-geometric theories of space-time [42]. Just as “value structure” solitons are ubiquitous in conventional CA models [32,74], “link structure” solitons might emerge in SDCA; physical particles would, in such a scheme, be viewed as geometrodynamical disturbances propagating within a dynamic lattice. Three SDCA-like theories of pregeometry have recently been proposed in which space-time is a self-organized emergent construct: Hillman [27], Nowotny and Requardt [55] and Wolfram [75].

Finally, we briefly comment on ostensible overlaps between SDCA and four other related fields of study: (1) *Lindenmeyer* (or L-) *systems*, (2) *graph grammars*, (3) *random graphs* (abbreviated, RG), and (4) *dynamic network analysis* (abbreviated, DNA). L-systems [57] are generalized CA systems in which the number of sites can grow with time, and consist of recursive rules for rewriting strings of symbols. If interpreted graphically, abstract symbol strings can be used to model growth processes of plants and evolving morphology of physical organisms. Graph grammars [20,35] apply formal language theory to networks, and consist of production rules that define the set of “valid structures” in a given graph language. The study of RG [15] was introduced by Erdos and Renyi in the late 1950s [16], and is a mathematical framework for exploring the general topological structures of computational systems and the behavior of certain random dynamical systems. Like SDCA, RG describes evolving graphs, but the dynamics are global and random. DNA [41,50] is an emerging field that fuses traditional social network theory with statistical analysis and modeling; part of its charter is to explore general properties of network generation and evolution.

While, conceptually speaking, there is a *prima facie* relationship between SDCA and all four fields of study, the elucidation of a more precise nature of the relationship between SDCA and these other systems awaits a future study. (The relationship appears particularly strong between SDCA and a generalized L-system called the *graph development system* (abbreviated, GDS), introduced by Doi [14], but not developed further since its original conception. Using incidence matrices to represent arbitrary topologies, GDS is essentially a grammar by which submatrices of the whole matrix are rewritten to describe topological changes. SDCA also formally falls under the broader rubrics of DNA and RG; however, there is no explicit reference to SDCA in the current literature of either field.)

### The Basic Model

Conventional CA are defined on fixed, and typically regular, lattices (one-dimensional lines, two-dimensional Euclidean or hexagonal grids, etc.), the sites of which are populated with discrete-valued dynamic elements ( $\equiv \sigma_i \in \{0, 1, \dots, k\}$ , where  $i$  labels a particular site on the lattice) that evolve according to local transition functions,  $f: \sigma_i \rightarrow \sigma'_i$ . We emphasize that the dynamics of conventional CA are confined to the temporal evolution of the  $\sigma_i$ s.

SDCA generalize conventional CA in two ways: (1) they relax the assumption that the underlying lattice is uniform, allowing the local site  $\leftrightarrow$  site connectivity pattern to vary throughout the lattice; and (2) they allow *both* the set  $\{\sigma_i\}$  and the lattice to evolve according to local transition rules. The most obvious – also the most dramatic – conceptual change this entails over the dynamics of conventional CA, is that the meaning of “local” itself changes as a function of how the SDCA system evolves: previously far separated sites may become neighbors; and sites that are local at time  $t$  may become far separated at some later time,  $t'$ .

To properly define SDCA, we first generalize regular lattices to mathematical graphs  $\equiv G()$  possessing arbitrary topology. Assuming  $G$  has  $N$  lattice sites, and that  $G$  is (for now) an *undirected* graph (meaning that none of  $G$ 's links carry directional information),  $G$  is completely defined by the  $N$ -by- $N$  *adjacency matrix*,  $\ell_{ij}$ :

$$\ell_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are linked;} \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Using the graph metric function,

$$D_{ij} = \text{Minimum}_{\text{Paths}, P_{ij}} [\# \text{links}, l_{rs} | \{r, s\} \in P_{ij}], \quad (2)$$

we can write a general  $r$ -neighborhood CA *value*-transition rule ' $f$ ' (which will from now on refer generically to as a  $\sigma$ -rule) in the form

$$\sigma_i^{t+1} = f \left[ \{\sigma_j^t | j \in S_r^G(i)\} \right], \quad (3)$$

where  $S_r^G(i) = \{j | D_{ij} \leq r\}$  is the radius- $r$  *graph sphere* about the site  $i$ . In words, the value of  $\sigma_i^{t+1}$  is some function,  $f$ , of the values  $\sigma_j^t$  in radius  $r$  graph sphere around the site  $i$ . With this distance measure,  $G$  becomes a *discrete metric space*. If  $G$  is a one-dimensional line, and  $r = 1$ , then  $S_r^G(i) = \{i-1, i, i+1\}$ ; i.e., it is equal to the conventional three site local neighborhood of elementary CA.

We now formally extend a conventional CA's dynamic arena – limited to the values  $\sigma_i^t \in \{0, 1, \dots, k-1\}$ ,

$i = 1, \dots, N$  – to one that includes the components of the underlying lattice's adjacency matrix:

$$\begin{cases} \sigma^{t+1} = F_\sigma[\{\sigma^t\}, \{\ell^t\}] \\ \ell^{t+1} = F_\ell[\{\sigma^t\}, \{\ell^t\}] \end{cases}, \quad (4)$$

where  $F_\sigma$  and  $F_\ell$  are some functions (to be defined explicitly below) that explicitly couple the changing *value* states and *geometries*. The complete system at time  $t$  is specified by the state-vector

$$|G\rangle_t = |\sigma_1^t, \dots, \sigma_N^t; \{\ell_{ij}^t\}\rangle. \quad (5)$$

The time-evolution of  $|G\rangle$  proceeds according to the following transition rules: (i)  $\sigma$ -rules of the general form given above and familiar from CA simulations and (ii)  $\ell$ -rules, which are divided into site *couplers*, linking previously unconnected vertices and site *decouplers*, which disconnect linked points. Because the topology can be altered only by either a deletion of *existing links* or an addition of links between pairs of vertices ' $i$ ' and ' $j$ ' with  $D_{ij} = 2$ , the dynamics is *strictly local*.

To be more precise, we first restrict the general  $\sigma$ -rule  $F_1$  to (maximally symmetric) *totalistic* (**T**) and *outer-totalistic* (**OT**) type. Since the underlying lattice is a fully dynamic object,  $|G\rangle$  will, in general, tend towards having a complex local geometry with an unspecified local *directionality*. The most general rules which can therefore be applied are those which are completely invariant under all rotation and reflection symmetry transformations on local neighborhoods. **T** (**OT**)  $\sigma$ -rules are then specified by listing particular *sums*  $\{\alpha\}$  (*outer-sums*  $\{\alpha_0\}, \{\alpha_1\}$  corresponding to center site values '0' and '1' respectively) for which the value of the center site becomes '1'. Formally,

$$\sigma_i^{t+1} = \phi_{\{\alpha\}} \left( \sum_j \ell_{ij}^t \sigma_j^t, \sigma_i^t \right), \quad (6)$$

where

$$\begin{aligned} \phi_{\{\alpha\}}(x, a) &= \begin{cases} \sum_\alpha \delta(x + a, \alpha) & \longleftrightarrow \mathbf{T} \\ a \sum_{\alpha_1} \delta(x, \alpha_1) + (1-a) \sum_{\alpha_0} \delta(x, \alpha_0) & \longleftrightarrow \mathbf{OT} \end{cases} \end{aligned} \quad (7)$$

and  $\delta(x, y)$  is the *Kronecker delta*. Note that  $\sum_j \ell_{ij}^t \sigma_j^t$  sums the values of all sites ' $j$ ' linked to ' $i$ ' at time ' $t$ '. The action on the state  $|G\rangle$  is represented by

$$\begin{aligned} \hat{\phi}_{\{\alpha\}}^i |\sigma\rangle_t &= |\sigma_1^t, \dots, \sigma_i^{t+1}, \dots, \sigma_N^t\rangle = \phi_{\{\alpha\}} \left( \sum_j \ell_{ij}^t \sigma_j^t, \sigma_i^t \right), \end{aligned} \quad (8)$$

where we distinguish the *operator*  $\widehat{\phi}^i$  acting on the global value state from the actual local transition *function*  $\phi$  which transforms each site value.

### Link Rules

Local geometry altering rules are constructed by direct analogy: for any two selected sites  $i$  and  $j$  we restrict attention to site values of vertices contained within a 1-*sphere* of either site; that is, to all  $k \in S_1(i, j) = S_1(i) \cup S_1(j)$ . Link operators, whose action on the state is represented by:

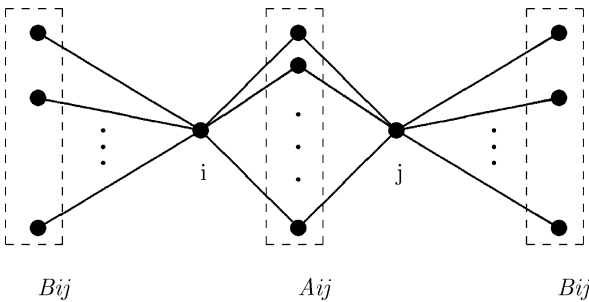
$$\begin{aligned} \text{decouplers: } \widehat{\psi}_{\{\beta\}}^{ij} \left| \ell_{ij}^t \right\rangle &= \left| \ell_{11}^t, \dots, \ell_{ij}^{t+1} = \psi^{ij}, \dots, \ell_{NN}^t \right\rangle \\ \text{couplers: } \widehat{\omega}_{\{\varepsilon\}}^{ij} \left| \ell_{ij}^t \right\rangle &= \left| \ell_{11}^t, \dots, \ell_{ij}^{t+1} = \omega^{ij}, \dots, \ell_{NN}^t \right\rangle, \end{aligned} \quad (9)$$

either *link* or *unlink* two sites ' $i$ ' and ' $j$ ' depending on whether the actual sum of values in  $S_1(i, j)$  matches any of those given in the  $\{\beta\}$  or  $\{\varepsilon\}$  lists, which completely define *decouplers* and *couplers*, respectively.

In order to construct classes of rules analogous to the two types of  $\sigma$ -rules defined above, we partition the local neighborhood into 3 disjoint sets (see Fig. 1):  $S_1(i, j) = V_{ij} \cup A_{ij} \cup B_{ij}$ , where

$$\begin{cases} V_{ij} = \{i, j\}, \\ A_{ij} = \{k | k \in C_1(i) \cap C_1(j)\}, \text{ where } C_1(i) = S_1(i) - \{i\}, \\ B_{ij} = S_1(i) \cup S_1(j) - V_{ij} - A_{ij}. \end{cases} \quad (10)$$

The action of link operators is then conveniently expressed as a function of the sums within the individ-



Structurally Dynamic Cellular Automata, Figure 1

**Neighborhood partitioning.** In the same way as *outer* sites can be considered separately for  $\sigma$ -transitions, we may, for topology transitions, distinguish between those sites belonging to *both*  $i$  and  $j$  ( $\in A_{ij}$ ) and those belonging to *one* of the two sites but not both ( $\in B_{ij}$ ). In this way we obtain the analogous *totalistic* (T), *outer-totalistic* (OT), and an additional type called *restricted totalistic* (RT)

partitions. Defining  $v_{ij} = \sigma_i + \sigma_j$ ,  $a_{ij} = \sum_{k \in A_{ij}} \sigma_k$ , and  $b_{ij} = \sum_{k \in B_{ij}} \sigma_k$ , we get *decouplers*,  $\psi_{\{\beta\}}^{ij} = \psi_{\{\beta\}}^{ij}(v_{ij}, a_{ij}, b_{ij})$ , where

$$\begin{aligned} \psi_{\{\beta\}}^{ij}(x, y, z) &= \begin{cases} \{1 - \sum_k \delta(x + y + z, \beta_k)\} \ell_{ij} & \leftrightarrow \mathbf{T} \\ \{1 - \sum_k \delta(x, \beta_{1,k}) \delta(y + z, \beta_{2,k})\} \ell_{ij} & \leftrightarrow \mathbf{OT} \\ \{1 - \sum_k \delta(x, \beta_{1,k}) \delta(y, \beta_{2,k}) \delta(z, \beta_{3,k})\} \ell_{ij} & \leftrightarrow \mathbf{RT} \end{cases} \end{aligned} \quad (11)$$

and *couplers*,  $\omega_{\{\varepsilon\}}^{ij} = \omega_{\{\varepsilon\}}^{ij}(v_{ij}, a_{ij}, b_{ij})$ , where

$$\begin{aligned} \omega_{\{\varepsilon\}}^{ij}(x, y, z) &= \begin{cases} \delta(D_{ij}, 2) \sum_k \delta(x + y + z, \varepsilon_k) & \leftrightarrow \mathbf{T} \\ \delta(D_{ij}, 2) \sum_k \delta(x, \varepsilon_{1,k}) \delta(y + z, \varepsilon_{2,k}) & \leftrightarrow \mathbf{OT} \\ \delta(D_{ij}, 2) \sum_k \delta(x, \varepsilon_{1,k}) \delta(y, \varepsilon_{2,k}) \delta(z, \varepsilon_{3,k}) & \leftrightarrow \mathbf{RT} \end{cases} \end{aligned} \quad (12)$$

In the above expressions, **RT** stands for *restricted totalistic* rules which maximally subdivide the local neighborhood. The inclusion of an  $\ell_{ij}$  in the expressions for  $\psi$  assures that only those sites already *linked* can be decoupled and the  $\delta(D_{ij}, 2)$  in the equations defining  $\omega$  are put in to make sure that only sites separated by distance = 2 may be dynamically coupled.

The three type-specific *sums* appearing above are indexed with the following conventions:

- T** rules are defined by the ' $k$ ' overall sums of values in  $S_1(i, j)$  for which the particular action is to be taken. For example, define ' $\psi$ ' by *unlinking* ' $i$ ' and ' $j$ ' if the total sum = 1 ( $= \beta_1$ ), 3 ( $= \beta_2$ ) or 5 ( $= \beta_3$ ). Equation (11) then states that  $\ell_{ij}^{n+1} = 0$  if and only if  $\ell_{ij}^n = 1$  and  $v_{ij}^n + a_{ij}^n + b_{ij}^n \in \{1, 3, 5\}$ .
- OT** rules are specified by giving ' $k$ ' 2-tuples  $(\beta_{1,k}, \beta_{2,k})$ , and  $(\varepsilon_{1,k}, \varepsilon_{2,k})$ , where  $\{1, k\}$  labels the sum ' $\sigma_i + \sigma_j$ ' and  $\{2, k\}$  labels the corresponding *outer sum*  $= \sum_{s \in S_1(i, j) - \{i, j\}} \sigma_s$ . For example, *link* ' $i$ ' and ' $j$ ' if  $\sigma_i + \sigma_j = 0$  and outer sum =  $\{3, 4\}$ , so that ' $\omega$ ' is defined by listing the two 2-tuples  $(\varepsilon_{1,1}, 0, \varepsilon_{2,1} = 3)$  and  $(\varepsilon_{1,2} = 0, \varepsilon_{2,2} = 4)$ .
- RT** rules are completely specified by giving the ' $k$ ' 3-tuples of values  $(x\sigma_i + \sigma_j, y = \text{sum in } A, z = \text{sum in } B)$ , for which the link operation between ' $i$ ' and ' $j$ ' is to be performed. For example, define ' $\psi$ ' by *unlinking* ' $i$ ' and ' $j$ ' for the following values of partitioned sums:  $(0, 0, 1)$ ,  $(0, 0, 2)$ ,  $(0, 1, 1)$ ,  $(1, 1, 1)$ ; we then have that  $(\beta_{1,1} = 0, \beta_{2,1} = 0, \beta_{3,1} = 1)$ ,  $(\beta_{1,2} = 0, \beta_{2,2} =$

$0, \beta_{3,2} = 2), (\beta_{1,3} = 0, \beta_{2,3} = 1, \beta_{3,3} = 1),$  and  $(\beta_{1,4} = 0, \beta_{2,4} = 1, \beta_{3,4} = 1).$

Global transition operators are obtained by applying individual  $\sigma$ - and  $\ell$ - operators to all *sites* and *site-pairs* in the graph  $G$ :

$$\begin{cases} \hat{\Phi}_{\{\alpha\}} |\sigma\rangle = \prod_i \hat{\phi}_{\{\alpha\}}^i |\sigma\rangle, \\ \hat{\Psi}_{\{\beta\}} |\ell\rangle = \prod_{n\langle ij\rangle} \hat{\psi}_{\{\beta\}}^{ij} |\ell\rangle, \\ \hat{\Omega}_{\{\varepsilon\}} |\ell\rangle = \prod_{nn\langle ij\rangle} \hat{\omega}_{\{\varepsilon\}}^{ij} |\ell\rangle, \end{cases} \quad (13)$$

where the products for  $\hat{\Psi}$  and  $\hat{\Omega}$  need to be taken only over *nearest* and *next nearest* pairs respectively. Given the full *value-topology* transition rule  $\Gamma$ , defined by

$$|G\rangle_{t+1} = (\hat{\Omega} \hat{\Psi} \hat{\Phi}) |G\rangle_t = \Gamma |G\rangle_t, \quad (14)$$

the fundamental problem is to understand the generic behavior of accessible graphs- $G$  emerging from all possible initial structures and value configurations. We emphasize that the lattice *fully* participates in the dynamics and that, in general, no embedding is implied – it is the abstract *connectivity* itself whose evolution we are attempting to trace.

### An Example

The application of the rather cumbersome expressions defining transition rules is in practice extremely straightforward, as we demonstrate with the following example: Consider a graph  $G$  defined as a  $(5 \times 5)$  lattice with some distribution of values  $\sigma = 1$  at time ' $t = 1$ ' (see Fig. 2). We are interested in one *global* update of the system

$|G\rangle_{t=1} \xrightarrow{\Gamma} |G\rangle_{t=2}$  with rules specified by

$$\begin{aligned} (\text{value}) \quad & \Phi_{\{\alpha\}} : \{\alpha\}_T = \{\alpha_1 = 1, \alpha_2 = 3, \alpha_3 = 5\}, \\ (\text{topology}) \quad & \Psi_{\{\beta\}} : \{\beta\}_{OT} = \left\{ \begin{array}{l} (\beta_{1,1} = 1, \beta_{2,1} = 3) \\ (\beta_{1,2} = 1, \beta_{2,2} = 4) \end{array} \right\}, \\ & \Omega_{\{\varepsilon\}} : \{\varepsilon\}_{OT} = \{\varepsilon_{1,1} = 1, \varepsilon_{2,1} = 3\}. \end{aligned} \quad (15)$$

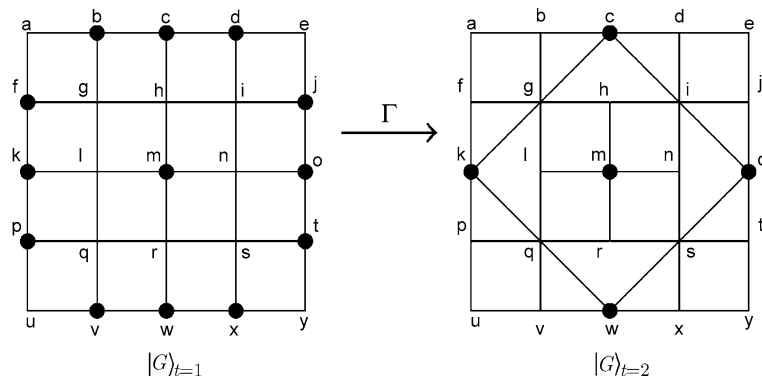
We evolve the system by systematically sweeping through all *sites*, *linked* pairs, and *next-nearest neighbors*:

1. *All Sites*: ... setting  $\sigma_i = 1$  only at those ' $i$ ' for which the sum of the values at ' $i$ ' and its neighbors is equal to ' $2$ ' at  $t = 1$ . By "neighbors" of any point ' $i$ ' we will always mean the set of vertices linked to ' $i$ ':  $(a, b)$ ,  $(h, m)$  and  $(x, y)$ , for example, are all neighbors at  $t = 1$ . Writing out a few value-changing terms explicitly, we find that

$$\begin{aligned} \sigma_c^{t=2} &= \phi(\sigma_b^{t=1} + \sigma_c^{t=1} + \sigma_d^{t=1} + \sigma_h^{t=1}) \\ &= \phi(3) = 1, \quad \text{and} \\ \sigma_b^{t=2} &= \phi(\sigma_a^{t=1} + \sigma_b^{t=1} + \sigma_c^{t=1} + \sigma_g^{t=1}) \\ &= \phi(2) = 0. \end{aligned} \quad (16)$$

2. *All linked pairs of sites ' $i$ ' and ' $j$ '*: ... removing those links only if the 2-tuple  $(\alpha, \beta) \in \{(1, 3), (1, 4)\}$ , where  $\alpha = \sigma_i + \sigma_j$  and ' $\beta$ ' is the sum of values of the neighbors of ' $i$ ' and ' $j$ ' at  $t = 1$ . For the points ' $c$ ' and ' $h$ ', for example, we have  $(\alpha, \beta) = (1, 3)$ , so that the link  $\ell_{ch}$  is no longer present in  $|G\rangle_{t=2}$ :

$$\begin{aligned} \ell_{ch}^{t=2} &= \psi(\sigma_c^{t=1} + \sigma_h^{t=1}, \sigma_b^{t=1} + \sigma_d^{t=1} + \sigma_g^{t=1} \\ &\quad + \sigma_i^{t=1} + \sigma_m^{t=1}) \ell_{ch}^{t=1} \\ &= \psi(1, 3)(1) = 0. \end{aligned} \quad (17)$$



Structurally Dynamic Cellular Automata, Figure 2

Sample dynamic update of a  $(5 \times 5)$  lattice from  $t = 1$  to  $t = 2$ , obeying a T-type  $\sigma$ -rule with  $\sigma \rightarrow \sigma'$  for local sums  $= 1, 3, 5$  (i. e.  $\alpha \in \{1, 3, 5\}$ ), and OT-type  $\ell$ -rules: (i) *link* for  $\{\varepsilon_{1,1} = 1, \varepsilon_{2,1} = 3\}$  and (ii) *unlink* for  $\{\beta_{1,1} = 1, \beta_{2,1} = 3\}$  and  $\{\beta_{1,2} = 1, \beta_{2,2} = 4\}$ . Solid sites indicate that  $\sigma = 1$



3. All next-nearest neighbors ‘i’ and ‘j’: ... linking them only if the 2-tuple  $(\alpha, \beta) = \{(1, 3)\}$ . By “next-nearest neighbor” we mean those pairs which are themselves unlinked but which share at least one other linked neighbor:  $(a, g)$ ,  $(h, r)$  and  $(w, y)$ , for example, are all next-nearest neighbors at  $t = 1$ . For ‘c’ and ‘g’ we find

$$\begin{aligned} \ell_{cg}^{t=2} &= \omega \left( \sigma_c^{t=1} + \sigma_g^{t=1}, \sigma_b^{t=1} + \sigma_d^{t=1} + \sigma_f^{t=1} \right. \\ &\quad \left. + \sigma_h^{t=1} + \sigma_l^{t=1} \right) \delta(D_{cg}, 2) \\ &= \omega(1, 3)(1) = 1. \end{aligned} \quad (18)$$

Notice that although  $\ell_{dn}^{t=1} = 0 \rightarrow \ell_{dn}^{t=2} = 1$ , it is hidden by overlap with the remaining links  $\ell_{di}^{t=2} = 1$  and  $\ell_{in}^{t=2} = 1$ . For this reason, not all link changes can always be observed directly in the following figures.

Other sites and links are updated in precisely the same manner. Had the link-rules been of **T**-type, only one sum would have to be considered: the sum of the values of the points in question along with their neighbors’ values. Had they been, instead, of **RT**-type, three sums would have to be considered: the sum of the values of the sites in question, the sum of the values of their common neighbors (neighborhood *A* in Fig. 1) and the sum of the values of the points that are neighbors of one of the considered points, but not of the other (neighborhood *B* in Fig. 1). The final state  $|G\rangle_{t=2}$  emerges after the above process has been applied concurrently to all pairs, neighbors and next-nearest neighbors in  $|G\rangle_{t=1}$ .

## Comments

We conclude this section by making a few important general comments:

*Comment 1.* As defined above,  $\Gamma$  consists of three operators acting simultaneously on the state  $|G\rangle$ . More generally, one may prescribe any of 10 possible time-orderings to the operators  $\Omega$ ,  $\Psi$  and  $\Phi$ . That is, specify certain intermediate state dependencies, so that, for example  $\Gamma_1|G\rangle \equiv (\Omega\Psi)(\Phi|G\rangle)$  would in general be expected to yield results different from, say,  $\Gamma_2|G\rangle \equiv \Omega(\Phi(\Psi|G\rangle))$ . While we will be solely concerned with the synchronous time ordering defined above, we do not expect the qualitative results to depend critically on this choice.

*Comment 2.* A given rule  $\Gamma$  is completely defined by the set of sums  $\{\alpha\}$ ,  $\{\beta\}$  and  $\{\varepsilon\}$ . Alternatively, we can conveniently summarize a chosen transition rule by its vector-code  $\vec{C} = (c[\phi], c[\psi], c[\omega])_{a,b}$ , where

$$\begin{aligned} c[\phi] &= \begin{cases} \sum_{\alpha} 2^{\alpha} & \leftrightarrow \mathbf{T} \\ \sum_{\alpha_0} 2^{2\alpha_0} + \sum_{\alpha_1} 2^{(2\alpha_1+1)} & \leftrightarrow \mathbf{OT} \end{cases} \\ c[\psi] &= \begin{cases} \sum_k 2^{\beta_k} & \leftrightarrow \mathbf{T} \\ \sum_k 2^{3\beta_{2,k} + \beta_{1,k}} & \leftrightarrow \mathbf{OT} \\ \sum_k 2^{3(\beta_{2,k} + a\beta_{3,k}) + \beta_{1,k}} & \leftrightarrow \mathbf{RT} \end{cases} \quad (19) \\ c[\omega] &= \begin{cases} \sum_k 2^{\varepsilon_k} & \leftrightarrow \mathbf{T} \\ \sum_k 2^{3\varepsilon_{2,k} + \varepsilon_{1,k}} & \leftrightarrow \mathbf{OT} \\ \sum_k 2^{3(\varepsilon_{2,k} + b\varepsilon_{3,k}) + \varepsilon_{1,k}} & \leftrightarrow \mathbf{RT} \end{cases} \end{aligned}$$

where  $a = \max\{\beta_{2,k}\} + 1$ ,  $b = \max\{\varepsilon_{2,k}\} + 1$ , and must be specified only for **RT**-type topology rules. The  $\Gamma$  appearing in the above example, therefore, can be summarized by  $c[\phi] = 42$ ,  $c[\psi] = 2^{3(3)+1} = 1024$  and  $c[\omega] = 2^{3(4)+1} + 2^{3(3)+1} = 9216$ . Note that ‘ $\Psi$ ’ and ‘ $\Omega$ ’ are chosen always to be of the same type.

*Comment 3.* Computer simulations of these systems require that some measures be taken to prevent possible memory overflows, such as would happen in cases either of pure coupling, where links are continually added and none deleted, or in isolated regions of a graph where for a few sites more neighbors are added than are allowed by memory. We thus introduce *working* link transition rules

$$\tilde{\psi}^{ij} \equiv \begin{cases} \psi^{ij} & \longleftrightarrow d'_i \text{ or } d'_j > \delta \equiv d_{\min} \\ 1 & \longleftrightarrow \text{else,} \end{cases} \quad (20)$$

$$\tilde{\omega}^{ij} \equiv \begin{cases} \omega^{ij} & \longleftrightarrow d'_i \text{ or } d'_j < \Delta \equiv d_{\max} \\ 0 & \longleftrightarrow \text{else,} \end{cases} \quad (21)$$

where  $d_i = \text{degree}(i)$  (i. e. number of neighbors of  $i$ ). In words: make a sweep of the lattice, temporarily storing the candidates to *add* and *delete* for each point. If, for any point  $i$ , the updated degree is greater than  $\delta$  then proceed with deleting the stored deletion-candidates, otherwise do not delete; similarly, provided that the updated degree is less than  $\Delta$  proceed with addition. Thus, it is sufficient that *one* of two points allow a dynamic link change between them for that change to be enacted. In the following, the complete constrained dynamics will be quoted as  $\vec{C}_{(a,b)}^{[\delta,\Delta]}$ . If constraints play no role in the actual evolution of specific examples, they will be left out of the definition.

*Comment 4.* Because each dynamic update involves three separate types of processing, the number of possible rules is extraordinarily large (see Table 1). Unlike pure  $\sigma$ -transitions, however, the fraction of the total number which yield interesting behavior (i. e. neither immediately explosive, where the number of links increases with-

Structurally Dynamic Cellular Automata, Table 1

Numbers of possible rules for each of the three types of transition rules.  $d = \text{maximum allowable degree}$  and  $a = \text{maximum sum to be used from partition } A_{ij}$ . Example: for  $d = 5$ , we have  $N_\phi = 4096$ ,  $N_\psi = 2^{24} \sim 2 \times 10^7$  and  $N_\omega = 2^{21} \sim 2 \times 10^6$ . We thus have  $N_T = N_\phi N_\psi N_\omega \sim 10^{17}$  possible type OT  $\Gamma$ 's

| Rule type | $\phi$     | $\psi$             | $\omega$           |
|-----------|------------|--------------------|--------------------|
| T         | $2^{d+1}$  | $2^{2d}$           | $2^{2d-1}$         |
| OT        | $2^{2d+2}$ | $2^{6(d-1)}$       | $2^{3(2d-3)}$      |
| RT        | —          | $2^{3(a+1)(2d-1)}$ | $2^{3(a+1)(2d+1)}$ |

out bound, nor immediately degenerative, where an initial graph rapidly dwindles to a few isolated links) appears to be manageably smaller.

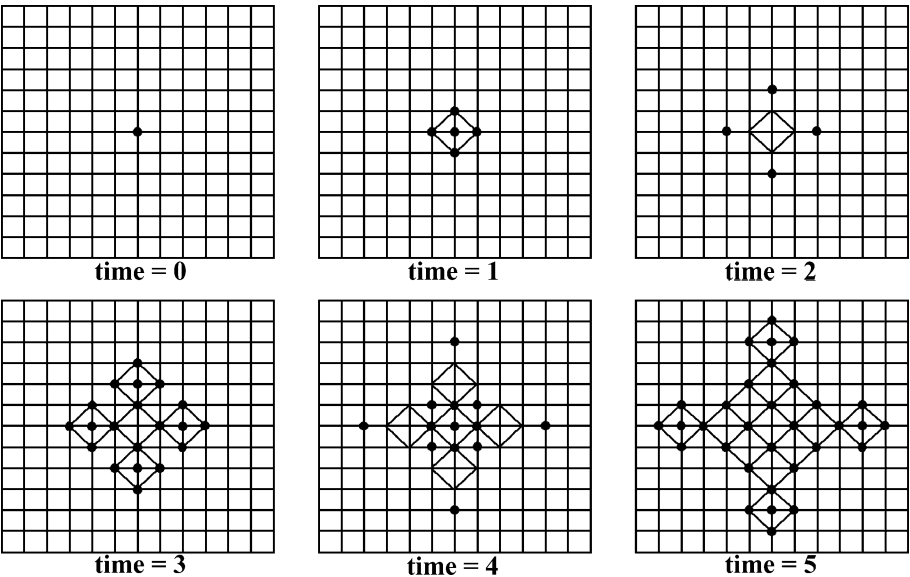
*Comment 5.* Although it is the intrinsic geometrical patterning whose generic behavioral properties we are trying to deduce, one may approach SDCA from an alternative point of view: maintain the emphasis on unraveling the value configurational behavior, and interpret the presence of  $[\Psi, \Omega]$  as background operators inducing nonlocal spatial connectivities. Whereas the systems defined above are completely abstract entities, in that locality is strictly defined by the link structure, the alternative scheme would be to embed the discrete networks in some specified manifold, and to study the effects of dynamically allocated non-local communication channels.

Emerging Patterns and Behaviors

Consider patterns that emerge from simple value seeds starting from ordered two dimensional Euclidean lattices. A single non-zero site may represent a small local disturbance that then propagates outward, restructuring the lattice. With appropriately chosen  $\Gamma$ 's one can induce a rich spectrum of different time evolutions only slightly perturbed by very few concurrent link changes to ones in which the initial geometry becomes radically altered. (The graphical representation of evolving one dimensional systems, in which link additions must be shown as *arcs* to avoid overlap with existing links, is needlessly confusing and is not considered.)

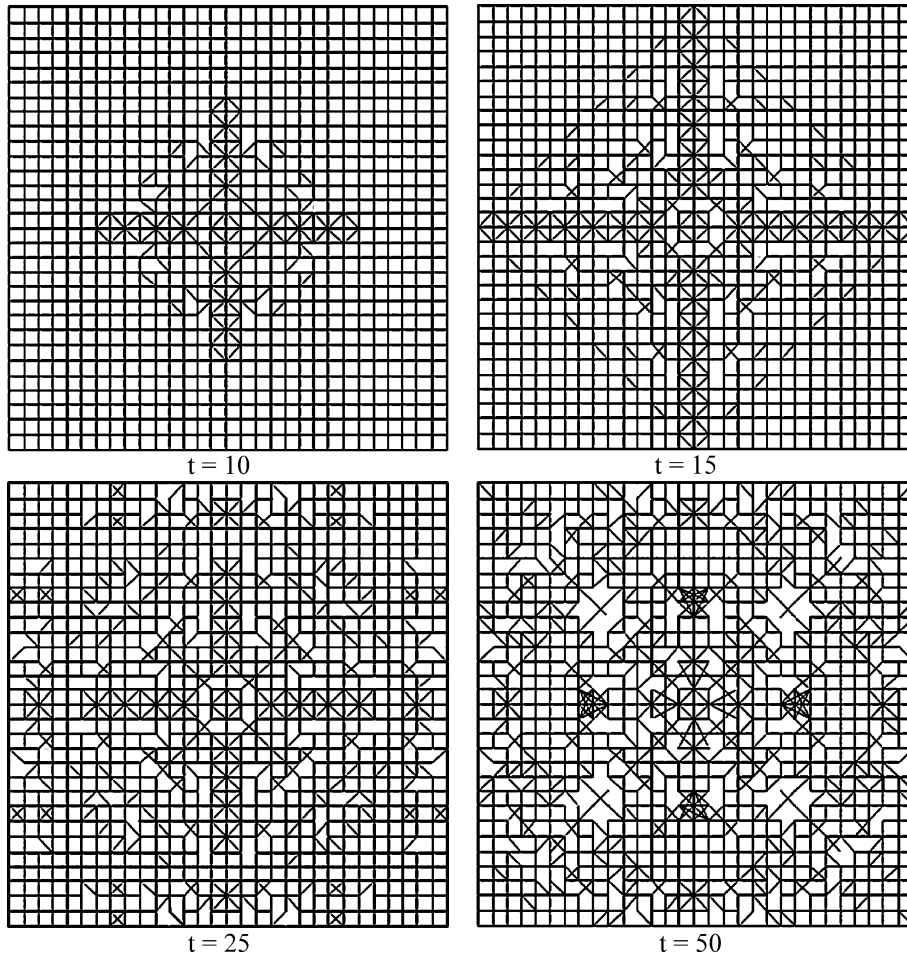
Figure 3 shows the first five iterations of a system starting from a four neighbor lattice with a single non-zero site at its center, the link structure is given explicitly and the solid circles represent sites with  $\sigma = 1$ . Notice how the link additions *follow* the emerging corrugated boundary surface of the value configuration. Remember that link additions are more than passive markers indicating particular correlations between local value configurations and structure; their presence directly influences all subsequent value development in their immediate vicinity.

Figure 4 (in which site values are suppressed for clarity) shows the continued development of this system. Though boundary effects begin to appear by  $t = 25$ , the



Structurally Dynamic Cellular Automata, Figure 3

First five iterations of an SDCA system starting from a 4-neighbor Euclidean lattice seeded with a single non-zero site at the center. The global transition rule  $\Gamma$  consists of T  $\sigma$ -rule and RT  $\ell$ -rules:  $\vec{C} = (26, 69648, 32904)_{[3,3]}$  (see text for rule definitions and code). Solid sites have  $\sigma = 1$



**Structurally Dynamic Cellular Automata, Figure 4**

Several further time frames in the structural evolution of the same system shown in the preceding figure. The values have been suppressed for clarity. The boundaries of the original lattice do not extend beyond the region shown so that the development is strictly confined to a  $31 \times 31$  graph

characteristic manner in which this particular  $\Gamma$  restructures the initial graph is clear:

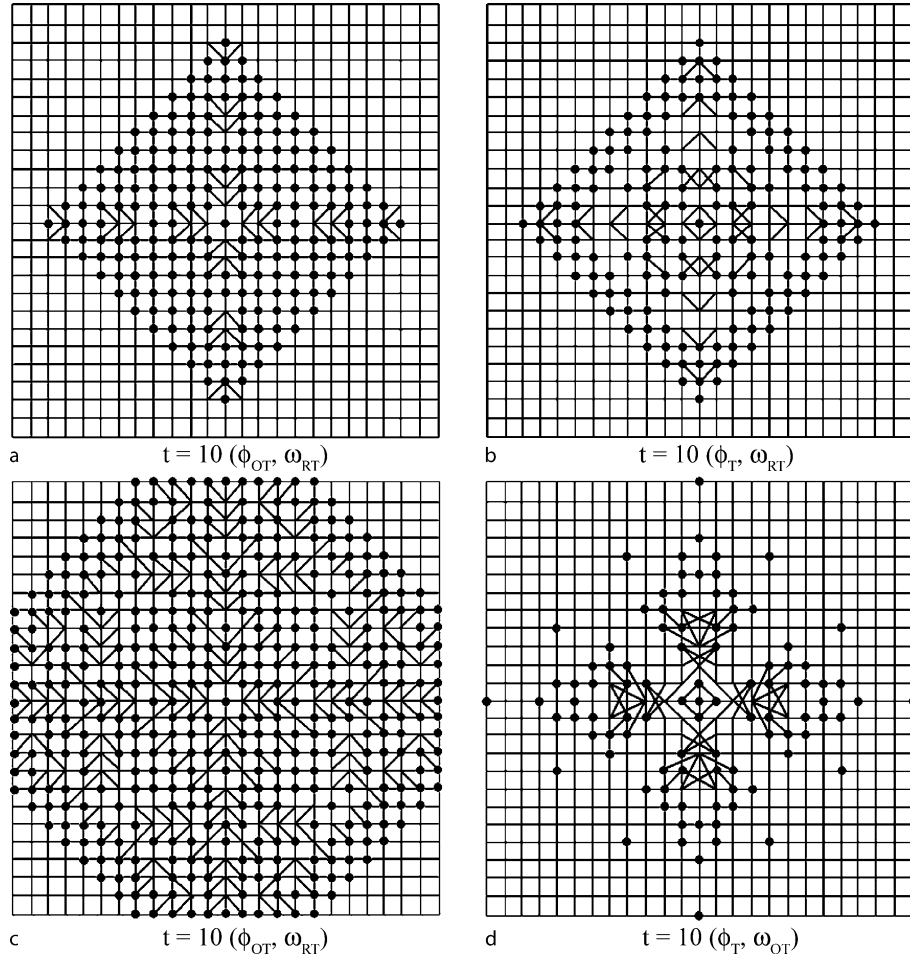
- There is a high degree of geometrical organization (the symmetry of the initial state is trivially preserved by the totally symmetric  $\Gamma$ ).
- The lattice remains connected.
- The distribution of link changes made throughout the lattice remains fairly uniform (i.e. there is an approximate uniformity in the probability of appearance of particular local value states which induce a structural change).
- Link-lengths do not get arbitrarily large.

The last point implies that for a system embedded in the plane, communication channels remain approximately

local. The global pattern emerges as a consequence of local ordering. On the other hand,  $\Gamma$ 's for which link-lengths get arbitrarily large are also easy to find.

Some other varieties of behavior are shown in Figs. 5 and 6. Figures 5a and b are representative of the class of  $\ell$ -rules that only mildly perturb the underlying lattice (and for which  $\sigma$  states do not differ much from their conventional CA cousins). Other rules, of course, may have a stronger effect on the lattice, giving rise to associated  $\sigma$  states bearing little or no resemblance to their conventional CA counterparts.

Figure 5c shows an example of a link rule that *accelerates* the outward propagation of the value configuration. Compare the diameter of this pattern to that in the earlier figures, both shown at equal times. The outwardly oriented



**Structurally Dynamic Cellular Automata, Figure 5**

Snapshot views of four typical developing states starting from a single non-zero site at the center of a 4-neighbor graph.  $\Gamma$ 's are as follows: **a** OT  $c[\phi] = 1022$  and RT coupler  $c[\omega] = 16, b = 1$ ; **b** T  $c[\phi] = 22$  and RT coupler  $c[\omega] = 32, b = 2$ ; **c** OT  $c[\phi] = 1022$  and RT coupler  $c[\omega] = 8, b = 1$ ; **d** T  $\sigma$ - and OT  $\ell$ -rules  $\bar{C} = (682, 19634061312, 133120)_{[2,8]}$

links that emerge from sites along the boundary surface become conduits by which non-zero values rapidly propagate. Had the underlying lattice topology been suppressed in this figure, and attention focused exclusively on the developing  $\sigma$  state, we could have interpreted the result as showing an effective increase in information propagation speed due to non-local connectivities (see comment 5 of the previous section).

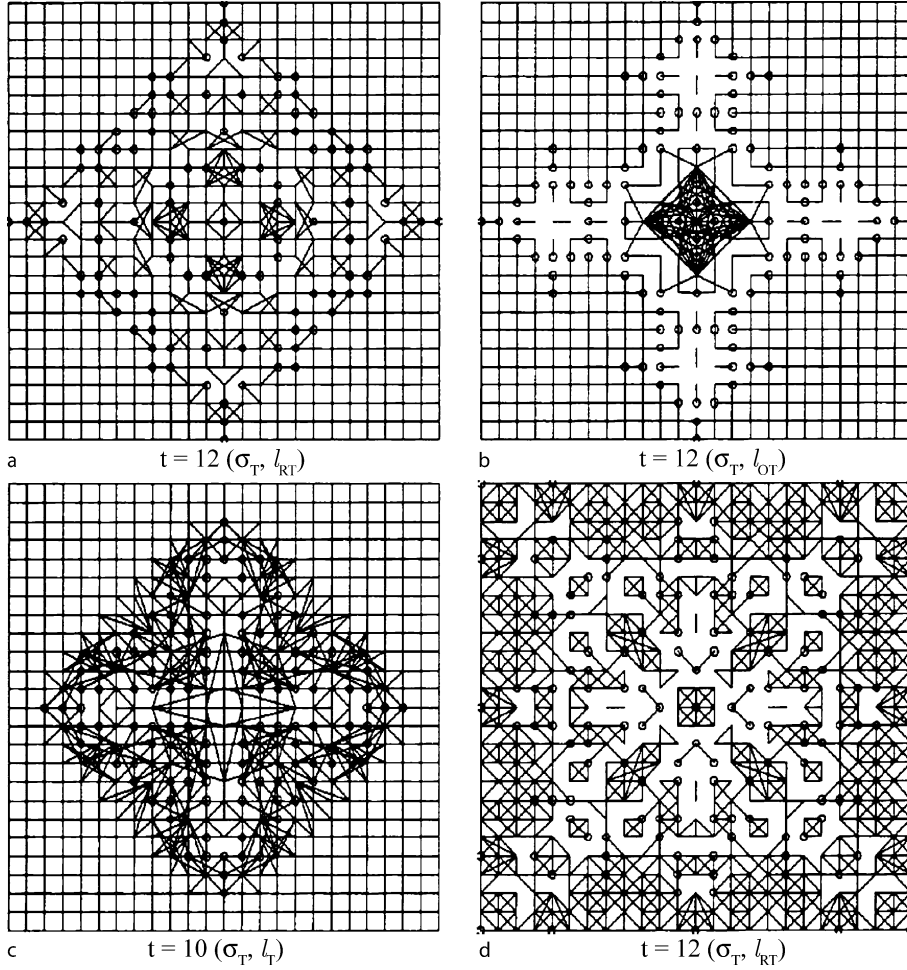
Figure 5d, on the other hand, gives an example in which the link dynamics *lags behind* the  $\sigma$  development. The boundary proceeds outward essentially unaffected by changes in geometry, which are themselves confined to the interior parts of the lattice (at least at this early stage of this system's development).

Figure 6 shows snapshot views of a few system undergoing a slightly more complex evolution. Figure 6b, for

example, shows a rule in which the outward  $\sigma$  propagation rapidly deletes most links from the original lattice but leaves a complex (though structurally stable) geometry at the origin of the initial disturbance. Figure 6c, on the other hand, shows a typical state of a system whose global connectivity becomes progressively more complicated.

A typical evolution starting from an initial state in which all sites are randomly assigned  $\sigma = 1$  with probability  $p = 1/2$  is shown in Fig. 7. Notice the rapid development of complex local connectivity patterns, the appearance of which points to a geometrical self-organization.

In general, structural behaviors emerging from random  $\sigma$ -states under typical  $\Gamma$ 's can be grouped into four basic classes (not to be confused with Wolfram's classification of elementary CA [74]):



**Structurally Dynamic Cellular Automata, Figure 6**

Four more examples of states emerging from simple seeds. Figure a, b, c start from 4-neighbor graphs and d from an 8-neighbor graph ( $\equiv$  4-neighbor with diagonals).  $\Gamma$ 's are as follows: **a**  $\sigma$ - and RT  $\ell$ -rules  $\tilde{C} = (42, 69648, 32904)_{[3,3]}$ ; **b**  $\sigma$ - and OT  $\ell$ -rules  $\tilde{C} = (42, 589952, 8192)_{[2,8]}$ ; **c**  $\sigma$ - and  $\ell$ -rules  $\tilde{C} = (42, 128, 4)_{[0,10]}$ ; **d**  $\mathbf{T} \mathbf{c}[\phi] = 682$  and RT  $\ell$ -rules defined explicitly by  $\Psi_{(104),(114),(124),(103),(113),(123)}$  and  $\Omega_{(111),(215)}$

- *Class-1*, in which initial graphs decay into structurally much simpler final states: most links are destroyed, and graphs  $\ell_{ij}^t$ , for sufficiently large  $t$ , consist essentially of a large number of small local subgraphs.
- *Class-2*, whose final states are characterized by periodic but globally connected geometries. SDCA typically arise in this class either because of a specific class-2  $\Phi$ s remaining unchanged by the coupling to the lattice or class-3  $\Phi$ s coupling with  $\{\Psi, \Omega\}$  in such a way as to induce a lattice structure that supports a periodic state.
- *Class-3*, consisting of SDCA that tend to grow in size and complexity, at least as measured by two basic metrics: the *average degree*,  $\langle \text{deg} \rangle \equiv (1/N) \cdot$

$\sum_i [|S_1(i)| - 1]$ , and *effective dimensionality*,  $D_{\text{effec}} \equiv \langle N_{\text{nn}} \rangle / \langle \text{deg} \rangle$ , where  $\langle N_{\text{nn}} \rangle$  is the average number of next-nearest neighbors. The values of both  $\langle \text{deg} \rangle$  and  $D_{\text{effec}}$  increase without bound for class-3 SDCA (unless an arbitrary upper constraint  $\Delta$  is imposed on  $\Gamma$ ). Because the  $\sigma$ -density responds to the changing local neighborhood structure, it is possible that what at first appears to be an explosive growth in fact eventually leads to a more sedate, if not static, behavior at some larger  $\langle \text{deg} \rangle \gg \Delta$ .  $\Phi$ s that yield  $\langle \sigma \rangle_t \sim \text{constant}$  over a range of  $\langle \text{deg} \rangle$  (such as the *sum modulo-2* rule; see below), when coupled with link rules that themselves become progressively less active with increasing  $\langle \text{deg} \rangle$ , may induce evolutions leading to only mild



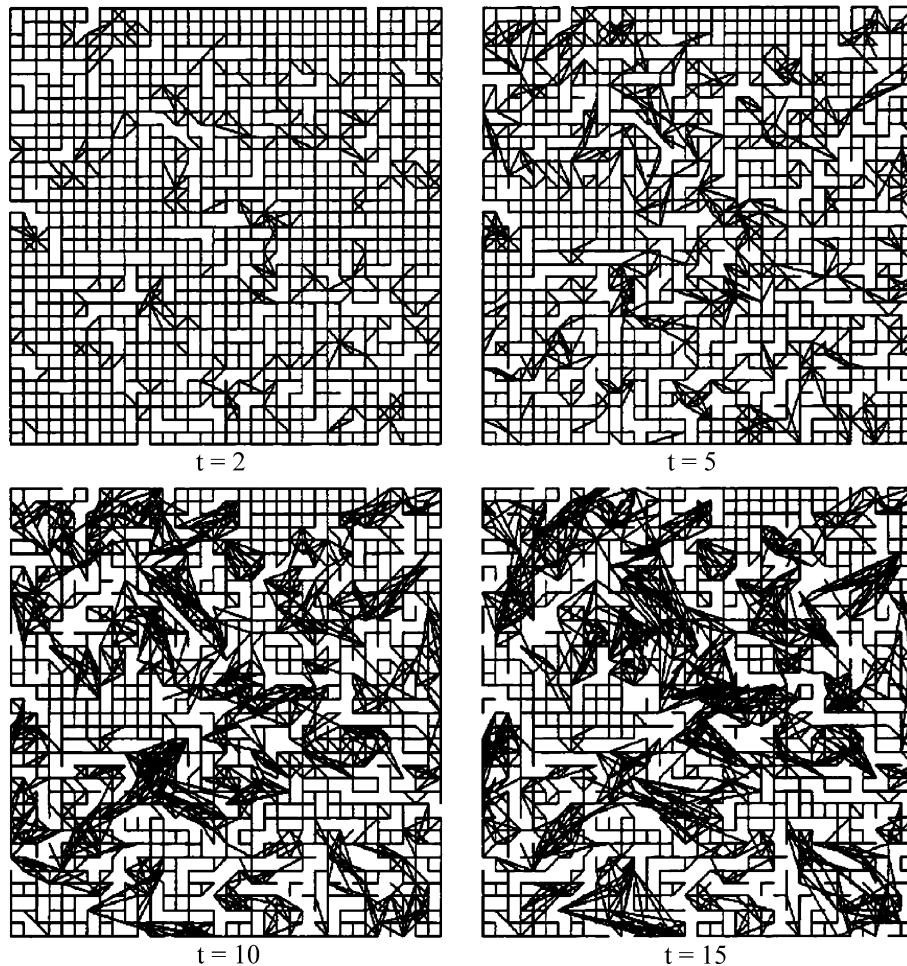
changes within specific ranges of the local structural parameters.

- *Class-4*, which is a provisional class (pending stronger evidence) that denotes a set of rules that yield open-ended  $\sigma$ - and  $\ell$  changes, but during which the value of  $D_{\text{effec}}$  remains roughly constant.  $\Psi$ s and  $\Omega$ s belonging to this class effectively induce a *structural equilibrium*: despite the fact that large numbers of link changes continue to be made, so that the detailed structure of the evolving graph continually changes, the average ratio of the number of next-nearest to nearest neighbors stays approximately constant over long periods of time. While there is evidence to suggest this class is *real*, simulations have unfortunately been run for too short a time and on graphs containing too

few sites to permit making any conclusive statements regarding the veracity of this class. Nonetheless, it is tempting to speculate that, for arbitrary values of  $D^*$ , there exists at least one set of SDCA rules for which  $D_{\text{effec}} \approx D^*$  (within a desired  $\epsilon > 0$ ) as the size of the graph  $N \rightarrow \infty$ . (*Pseudo* class-4 behavior, of course, can always be artificially induced either by imposing severe  $[\delta, \Delta = \delta]$  constraints, or, as must typically be done for category-3  $\Gamma$ s, by deliberately impeding growth with some threshold  $\Delta$ .)

### Statistical Measures

As evidenced by Fig. 7, it is already nontrivial to meaningfully visualize the short-time evolution of (initially) *regu-*



**Structurally Dynamic Cellular Automata, Figure 7**

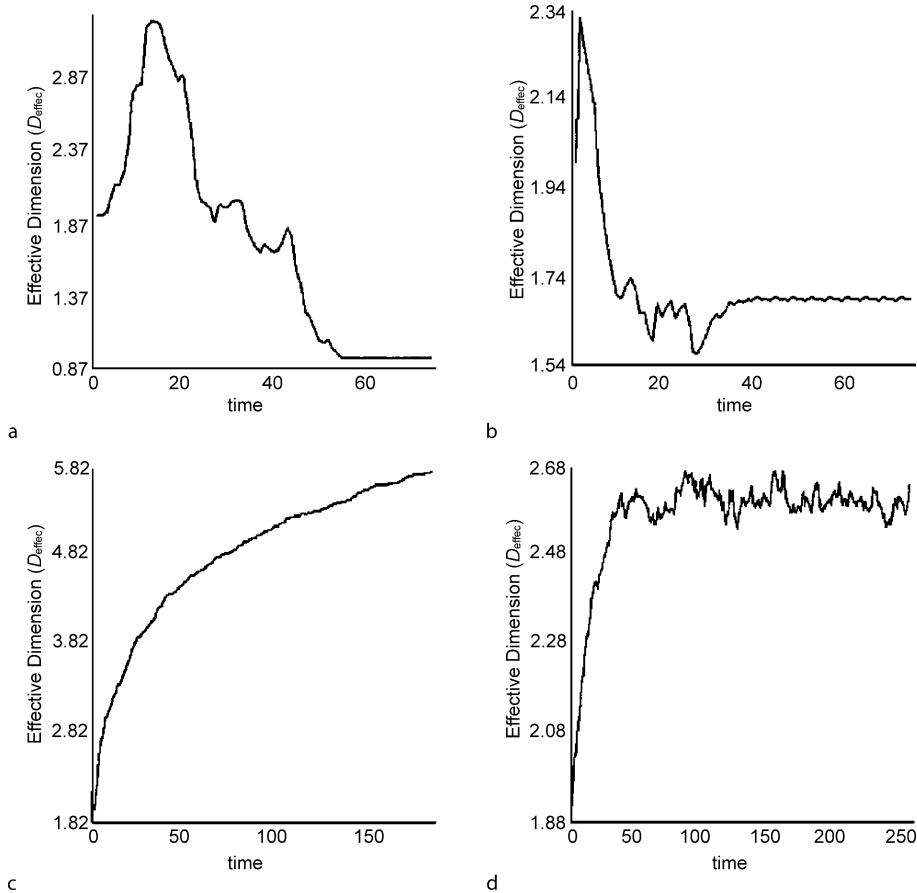
Evolution of a  $35 \times 35$  lattice, with randomly seeded sites. The development proceeds according to  $T \sigma$ - and  $OT \ell$ -rules defined by code  $\vec{C} = (84, 36864, 2048)$ . The constraints are  $[\delta = 0, \Delta = 10]$ . The appearance of localized substructures is evidence of a geometrical self-organization

lar lattices that start with random initial value state. Visualizing the long-term dynamics of systems that start from a completely random state is even more difficult (although graph visualization algorithms may help). However, even in cases for which a direct visual inspection of the dynamics reveals little, one can always indirectly keep abreast of a given system's properties by monitoring its core structural and behavioral measures (a more detailed account is given in [31]).

Site value measures include the average density of sites with value  $\sigma = 1$ ,  $\langle \sigma \rangle_t \sim (1/N) \sum_{i=1}^N \sigma_i^t$ ; the local value correlation,  $C^t \equiv \langle \sigma_i^t \cdot \sigma_j^t \rangle - (\rho^t)^2$ , where  $\langle \sigma_i^t \cdot \sigma_j^t \rangle$  is averaged over all pairs  $i$  and  $j$  with  $\ell_{ij} = 1$ ; the fraction of sites whose value *changes* during one step of the evolution,  $\Delta_t \equiv (1/N) \sum_{i=1}^N \{\sigma_i^{t-1} \oplus \sigma_i^t\}$ , where  $\oplus_2$  is a sum modulo-2.

Geometry measures include the *average degree*,  $\langle \deg \rangle$ ; the *average number of next-nearest neighbors*,  $\langle N_{nn} \rangle_t \equiv (1/N) \sum_i [|S_2(i)| - |S_1(i)| - 1]$ ; and  $D_{\text{effec}}$ . A measure of how the actual size of local neighborhoods changes with time may be obtained by embedding graphs into the two-dimensional plane and calculating the average pathlength at time  $t$ . Of course, global features that describe all complex networks – such as *connectivity*, *density*, *clustering*, and *path lengths* (2,6), are applicable to SDCA as well.

Link changes may be monitored by keeping track of (1) the total number of *link changes* (allowed under prescribed constraint conditions),  $\Delta_t^{(l)} \equiv (1/2) \sum_{i=1}^N \cdot \sum_{j=1}^N \{l_{ij}^t \oplus l_{ij}^{t-1}\}$ ; (2) the *constraint influence*,  $f_l \equiv \Delta_t^{(l)} / N_t^{(l)}$ , where  $N_t^{(l)}$  is the total number of link changes that would have occurred in the absence of constraints ( $f_l = 1$  indicates that the evolution is *pure*,



Structurally Dynamic Cellular Automata, Figure 8

Time development of the effective dimensionality  $D_{\text{effec}}$  for each of the four categories of behavior (see text): **a** T type  $I'$  defined by  $\tilde{C} = (42, 128, 4)$ ; **b** T  $\sigma$ - and OT  $\ell$ -rules  $\tilde{C} = (64, 9216, 1024)$ ; **c** T  $\sigma$ - and OT  $\ell$ -rules  $\tilde{C} = (682, 512, 512)_{[0,10]}$ ; **d** T  $\sigma$ - and RT  $\ell$ -rules defined explicitly by  $\tilde{\beta} \in \{(011), (110), (121), (233), (243)\}$  and  $\tilde{\epsilon} \in \{(120), (010), (021), (224)\}$

meaning it is unaffected by constraints;  $f_l \sim \text{small}$  suggests that the imposed constraint window  $[\delta, \Delta]$  has resulted in observed structures that are *impure*); (3) the link *creation-* and link *deletion-ratios*,  $f_C \equiv N_C/\Delta_t^{(l)}$  and  $f_D \equiv N_D/\Delta_t^{(l)}$ , where  $N_C$  and  $N_D$  are the numbers of link created and destroyed, respectively; (4) the *activity levels*,  $\gamma_C^t \equiv N_C/N_{nn}^{t-1}$  and  $\gamma_D^t \equiv N_D/N_l^{t-1}$  (where  $N_l^{t-1}$  is the number of links at time  $t-1$ ), which give the number of dynamic alterations relative to the corresponding spaces from which the candidates for alteration are selected; and (5) the *link evolution index*,  $\gamma_L^n \equiv (1/N_l^{t=0}) \sum_i \sum_j \{l_{ij}^{t=0} \oplus_2 l_{ij}^n\}$ , which gives the fraction of the initial lattice remaining after  $n$  iterations.

Figure 8 shows time series plots of  $D_{\text{effec}}$  for rules in each of the four behavioral classes defined above. The initial structure in each case is  $35 \times 35$  4-neighbor Euclidean lattice, so that  $D_{\text{effec}}^{t=0} \sim 2$ . Figure 8a gives an example of class-1 behavior, in which a short period of initial growth is followed by a decay into mostly disconnected clusters. The final state is characterized by  $\langle \text{deg} \rangle < 1$ , and is stable. Figure 8b shows a system that starts from the same initial state as in Fig. 8a but whose  $\Gamma$  leads to a periodic geometry. Just the right number of links have been deleted to permit regions with isolated activity to emerge.

Figure 8c shows class-3 behavior in which  $D_{\text{effec}}$  steadily increases. The apparent leveling off seen toward the end of the run is due both to a decreased overall activity level and the increasingly effect of the  $\Delta = 10$  constraint. The system in Fig. 8d exhibits class-4 behavior, characterized by an ongoing structural development within a relatively narrow interval of values of  $D_{\text{effec}}$ . Note that the structural changes here are essentially *pure*, and are not merely artifacts of any imposed constraints. Ilachinski (3) explores a wide range of emergent behaviors across all four classes, and examines the qualitative relationship between emergent behavior and initial  $\sigma$ - and  $\ell$ -seeding.

### Phase Plots

While it is of obvious interest to systematically explore every possible combination of  $\beta$ 's and  $\epsilon$ 's that define  $\ell$ -rules, Table 1 unfortunately suggests that the resulting rule space is simply too large. Nonetheless, we can learn much even by focusing our attention on a small subset of the complete rule space, keeping  $\Phi$ , the initial  $\sigma$ -seeding, and all other factors constant. Specifically, consider the subset of all possible  $\ell$ -rules that consists of **OT** link rules consisting of a single *coupler*,  $\omega$ , and a single *decoupler*,  $\psi$ . Moreover, let  $\phi \equiv \oplus_2$  (i. e., sum *modulo-2* rule), demand that only pairs of  $\sigma = 0$  sites be considered for a link change, and

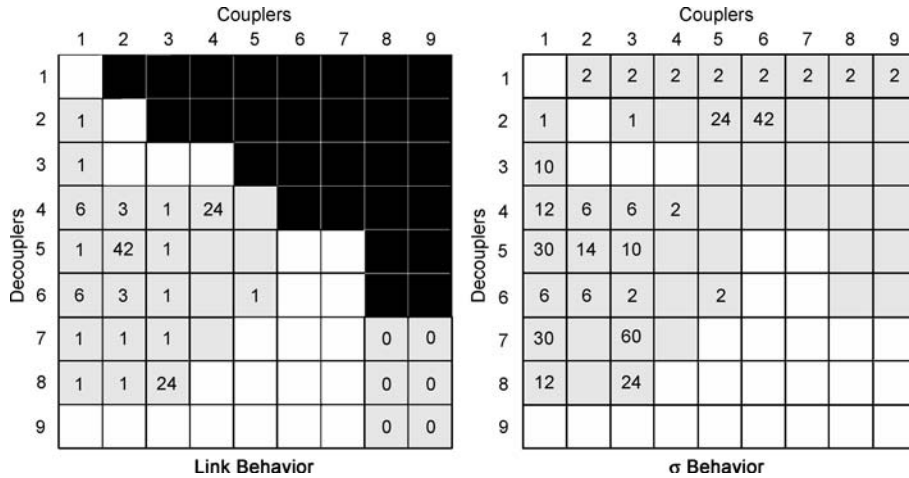
consider  $\ell$ -rules belonging to the following set:

$$\begin{aligned} \text{decouplers: } & \{\beta_{1,1} = 0, \beta_{2,1} = \mu\}, \quad 1 \leq \mu \leq 0, \\ \text{couplers: } & \{\epsilon_{1,1} = 0, \epsilon_{2,1} = \nu\}, \quad 1 \leq \nu \leq 0. \end{aligned} \quad (22)$$

Figure 9 summarizes the behavior of a four neighbor,  $25 \times 25$  lattice with periodic boundary conditions, starting from an initial  $\sigma$ -seed consisting of a single nonzero site. Four basic kinds of structural behaviors emerge:

1. *Static state*: this trivially occurs when the link rules are unable to take effect; namely, when  $\mu \geq 7$  and  $\nu \geq 8$ .
2. *Rapid growth*: for an entire range of  $\mu$  and  $\nu$ , the average number of neighbors for each site of the lattice increases rapidly for 20–30 iterations. This number would likely continue to increase, were it not for the constraint conditions ( $\equiv [0, 10]$ ). The “final state” is neither stable nor periodic. One sometimes also sees *delayed growth* in this class of behavior, in which case the link structure is initially relatively quiescent (and the behavior of the system as a whole mimics that of a conventional CA). As coupler rules are triggered by specific  $\sigma$  states, the average degree of the lattice rapidly increases (at least until the constraint conditions take effect).
3. *Spontaneous decay*: when decouplers are stronger than couplers, the average degree typically decreases. If this occurs too rapidly, the structure surrounding the single nonzero valued site may become isolated from other parts of the lattice. If a few non-zero values do not leak out into the outlying regions, link changes remain confined to the central subgraph, leading to either rapid stability or periodicity.
4. *Initial growth, followed by periodicity*: this is the least common behavior, and requires a delicate balance between coupler and decoupler rules.

It is interesting to compare these results with those obtained from a random  $\sigma$  seed. In this case, the sharp divisions between characteristic behaviors disappear, and there is a pronounced increase in the number of links for all  $\mu$  and  $\nu$ . However, the inclusion of an additional decoupler, may induce decay and periodicity. For example, consider the same initial lattice and  $\Phi$  as used in Fig. 9, fix two **OT**  $\ell$ -rules  $\Psi_{(0,5)}$  and  $\Omega_{(0,1)}$ , and add the decoupler  $\Psi_{(0,\mu)}: \{\beta_{1,1} = 0, \beta_{2,1} = \mu\}$ ,  $1 \leq \mu \leq 9$ . Surveying the emergent behaviors for this range of  $\mu$ 's, one now finds decaying lattices for  $\mu \geq 2$ . In each case, the initial graph succumbs to periodicity following a transient of between 50 and 100 iterations. The evolving lattice is also more prone to break up into small disconnected subgraphs.



**Structurally Dynamic Cellular Automata, Figure 9**

Phase plot that summarizes behavior of a four neighbor,  $25 \times 25$  lattice with periodic boundary conditions, starting from an initial  $\sigma$ -seed consisting of a single nonzero site.  $\Gamma$  is defined by the sum modulo-2  $\sigma$ -rule and  $\ell$ -rules of the form: *decouplers*– $\{\beta_{1,1} = 0, \beta_{2,1} = \mu\}$ , *couplers*– $\{\epsilon_{1,1} = 0, \epsilon_{2,1} = \nu\}$ . Grey areas in both plots denote *periodic states*. White areas denote *growth* in the plot for link behavior, and a *nonperiodic state* for  $\sigma$ -behavior. The black area that appears in the link-behavior plot denotes *decay*. Numbers that appear in individual boxes denote period lengths

Although, just as in conventional CA, small changes to  $\ell$ -rules can lead to large differences in emergent behavior, they generally appear to do so in a more predictable and patterned manner. Of course, particular classes of  $\Gamma$  may induce more complex phase plots; for example, isolated pockets of anomalous (and rapidly shifting) behavior may appear within larger surrounding regions undergoing otherwise mutually consistent and slowly changing dynamics. A better sense of the space of possible emergent behaviors, along with a deeper understanding of the relationship between  $\Phi$  and  $\ell$ -rules, awaits a future study.

### SDCA as Models of Computation

The basic SDCA model, as outlined above (which we will denote as  $\text{SDCA}_0$  to avoid possible confusion with the hierarchy of related SDCA models introduced in this section), was modified and generalized by Majercik [39] into a form more suitable for addressing its formal *computational capabilities* rather than as an exploratory toolkit for describing physical processes (which is the primary reason for which  $\text{SDCA}_0$  were first conceived). Motivated primarily by finding models of human brain function (for which one intuitively expects nonlocal neural connections to play a fundamental role in the rewiring of neural tissue), Majercik shows that suitably generalized SDCA are not only capable of universal computation, but actually represent a more efficient class of computational models than conventional CA. Majercik also reports an SDCA that can

solve the *firing squad* problem in  $O(\log t)$  time (i. e., exponentially faster than the  $O(t)$  in conventional CA), and a class of CA-*universal* SDCA models that can simulate any conventional CA with a speedup factor of two. (The *firing squad* problem [46] consists of finding a rule for which all sites in a CA evolve into a special state after the exactly the same number of steps.)

Majercik proceeds by first identifying five properties of  $\text{SDCA}_0$  that, while reasonable from a physical modeling standpoint, make it difficult to rigorously formulate and prove theorems:

1. *Finiteness*: The requirement that  $\text{SDCA}_0$  be strictly finite, both in time and space, is obviously necessary for computer experiments, but is unnecessarily restrictive for general theorem proving. Likewise, the assumption that the sets  $\alpha$ ,  $\beta$  and  $\epsilon$  must be finite is questioned.
2. *Bidirectionality*: While  $\text{SDCA}_0$  are defined with symmetric links, an obvious generalization that makes the basic model more readily applicable to neural dynamics (among other kinds of physical and biological systems) is to allow for unidirectional links.
3. *Link-rule Asymmetry*: While  $\text{SDCA}_0$ 's link decoupler function (Eq. (11)) contains the factor  $\ell_{ij}$  to explicitly prevent the system from inadvertently linking two unlinked sites,  $\text{SDCA}_0$  does not include an analogous term for the coupler function (that is, a term to prevent an evolving system from inadvertently unlinking two linked sites).

4. *Inconsistency*: While  $\sigma$ -rules effectively ignore site positions, all three types of link rules assume that the various neighborhoods surrounding individual sites ( $A_{ij}$ ,  $B_{ij}$ , and  $C_{ij} \equiv \{k | D_{ik} = 1 \oplus D_{jk} = 1\}$ , where ' $\oplus$ ' denotes *exclusive or*) are all *recognized as such* by the dynamics. That is, the link rules effectively “know” the positions of a site’s neighbors, while  $\sigma$ -rules possess no such information.
5. *Small Rule Set*: The class of  $\sigma$ - and  $\ell$ -rules used by SDCA<sub>0</sub> may be generalized to include a far broader class of transition functions.

On the basis of these observations, Majercik [39] introduces a set of three core models to define a hierarchy of eight alternative SDCA computational systems,  $\{\text{SDCA}^{(1)}, \text{SDCA}^{(2)}, \dots, \text{SDCA}^{(8)}\}$ . The three core models are (1) the *relative location* model ( $= M_R$ ), (2) the *labeled links* model ( $= M_L$ ), and (3) the *symmetric links* model ( $= M_S$ ). They differ only in the degree to which their  $\sigma$ - and  $\ell$ -transition functions depend on *specific* sites. For example,  $M_R$ ’s transition functions depend on the state and exact relative position of each neighbor (and therefore “knows” the exact source of any state in a local neighborhood). In  $M_L$ , links are labeled and the transition functions know both neighbor states and the label of the links to given neighbors, but the exact neighbor locations remain unspecified. Finally, in  $M_S$ , it is assumed that no information about the source of the neighborhood states exists, and transition functions only know the number of neighbors in a particular state.

Each of the three core models may be defined in two versions: an *unbounded links* (abbreviated, UL) version, in which the number of neighbors a given site can have is unbounded, and a *bounded links* (abbreviated, BL) version, in which an explicit upper limit is imposed. In addition, there is also one *finite labels* version of  $M_L$ . Majercik imposes certain mild conditions on the local transition functions; for example, that local neighborhoods always remain strictly finite,  $\sigma$ -rules leave quiescent neighborhoods alone, and that links between sites with quiescent neighborhood remain unaltered.

**Relative Location SDCA model** In the *Relative Location* model, the transition functions all have access to the exact relative *location* and *state* of each neighbor site. Define a *neighbor* of site  $i$ ,  $n_i \in S \times Z^d$ , as a pair that specifies the state (by a single label) and relative location of the neighboring site (as a  $d$ -tuple of coordinates). Let  $W = S \times Z^d$  be the set of all possible neighbors, and  $\mathcal{F}_W$  (called the *neighborhood function*) be the set of all possible finite, nonempty, partial functions that map  $Z^d$  to  $W$ . The local

state transition function  $\sigma: \mathcal{F} \rightarrow W$  maps neighborhood functions to the state set of SDCA. The local link transition function  $\lambda: \mathcal{F} \times \mathcal{F} \times \{0, 1, 2\} \rightarrow \{0, 1\}$  maps pairs of neighborhood functions (that define the neighborhoods of two sites,  $i$  and  $j$  and a number that specifies the status of the link between  $i$  and  $j$ : value *zero* meaning that  $i$  and  $j$  are neither direct neighbors nor next-nearest neighbors; value *one* meaning that  $i$  and  $j$  are immediate neighbors; and value *two* meaning  $i$  and  $j$  are next-nearest neighbors) to one of two link states: *zero*, meaning no link between  $i$  and  $j$ , and *one*, meaning a link exists.

**Labeled Links SDCA model** The *Labeled Links* model removes from  $M_R$ ’s transition functions any dependency on the exact relative location of a site’s neighbors, but allows the links to still be *labeled* so that the transitions functions can distinguish one link from another. This ability to “label” links paves the way for us to define SDCA with unidirectional links, since the labels can be used to distinguish between the *input* and *output* links to a site. Consider, for example, the UL-version of  $M_L$ . Labeling the links by natural numbers,  $\mathcal{N}$ , we define a neighbor of site  $i$ , as a pair  $(q, n)$ , where  $q \in S$  labels the state of the neighboring site, and  $n \in \mathcal{N}$  labels the link between  $i$  and its neighbor. Site  $i$  is defined as the direct neighbor linked via the 0th link, and the set of all possible neighbors,  $W = S \times \mathcal{N}$ . As for  $M_R$ ,  $\mathcal{F}_W$  is the set of neighborhood functions that map  $Z^d$  to  $W$ , the local state transition function  $\sigma: \mathcal{F} \rightarrow S$  maps neighborhood functions to states in  $S$ , and the local link transition function  $\lambda: \mathcal{F} \times \mathcal{F} \times \{0, 1, 2\} \rightarrow \{0, 1\}$  maps pairs of neighborhood functions and a number to either the values *zero* (unlinked) or *one* (linked).

**Symmetric Links SDCA model** The *Symmetric Links* model imposes the strictest constraint of all by doing away with all means by which the local transition functions may distinguish different neighborhood orientations. Consider the unbounded link version of  $M_S$ . Assume the SDCA has a total of  $n$  states, and let  $S = \{1, 2, \dots, n\}$ . Let  $\vec{n}_i \in N^n$  be an  $n$ -dimensional vector such that  $(n_i)_k$  is equal to the number of site  $i$ ’s neighbors in state  $k$ . Then the local state transition function  $\sigma: N^n \rightarrow S$  maps vectors in  $N^n$  to states in  $S$ , and the local link transition function  $\lambda: N^n \times \{0, 1, 2\} \rightarrow \{0, 1\}$  maps a vector in  $N^n$  and a link status label to either the values *zero* (unlinked) or *one* (linked).  $M_S$  can also be modified slightly to allow the local transition functions to retain knowledge of the state of site  $i$ : simply let  $\sigma: S \times N^n \rightarrow S$  map the pair consisting of the *state* of site  $i$  and a vector that defines the distribution of states among  $i$ ’s immediate neighbors (excluding  $i$ ). The local link function likewise assumes a simi-



lar form  $\lambda: S^2 \times N^n \times \{0, 1, 2\} \rightarrow \{0, 1\}$ , where the first component of the 3-tuple input to  $\lambda$  is a pair that defines the states of the two sites to which the link function is being applied.

### SDCA as CA Simulators

What does it mean to say that one dynamical system *simulates* another? Heuristically, it means that, for certain initial states, one system behaves just like another [32,74]. Suppose we have two CA systems – CA and CA' – defined by rules  $\phi$  and  $\phi'$ , and initial states  $\vec{\sigma} \in \Sigma$  and  $\vec{\sigma}' \in \Sigma$ , respectively. Then, loosely speaking,  $T$  iterations of CA are said to be “simulated” by  $nT$  ( $n \geq 1$ ) iterations of CA', provided there exists some invertible function,  $f: \Sigma \rightarrow \Sigma$ , by which  $\vec{\sigma}'$  is replaced by  $f(\vec{\sigma})$ . Simulation is a transitive relationship: if system B simulates system A, and another system C simulates B, then C also simulates A.

For example, a single site with a particular value in CA may be simulated by a fixed *block of sites* in CA'. After  $n$  steps, the blocks in CA' evolve to exactly the same final state as the single time-step evolution of individual sites in CA. As a concrete example, consider the elementary (one-dimensional, binary valued, conventional CA) rules  $\phi_{18}$  and  $\phi_{90}$ :

|               | 111 | 110 | 101 | 100 | 011 | 010 | 001 | 000 |
|---------------|-----|-----|-----|-----|-----|-----|-----|-----|
|               | ↓   | ↓   | ↓   | ↓   | ↓   | ↓   | ↓   | ↓   |
| $\phi_{18}$ : | 0   | 0   | 0   | 1   | 0   | 0   | 1   | 0   |
| $\phi_{90}$ : | 0   | 1   | 0   | 1   | 1   | 0   | 1   | 0   |

Provided that *two time steps* under  $\phi_{18}$  are carried out for every time step of rule  $\phi_{90}$ , it is easy to show that under the block transforms  $0 \rightarrow f(0) = 00$  and  $1 \rightarrow f(1) = 10$ , the evolution of arbitrary starting configurations under  $\phi_{90}$  is reproduced – or *simulated* – by  $\phi_{18}$ . For example, the global state  $\vec{\sigma} = '0011000'$  – which evolves into  $\phi_{90}(\vec{\sigma}) = '0111100'$  under  $\phi_{90}$  – yields the same state (after it is block-transformed) that results from two iterations of  $\phi_{18}$  applied to  $\phi_{90}$ 's block-transformed initial state,  $f(\vec{\sigma}) = '00001010000000'$ :

$$\phi_{18}[\phi_{18}[f(\vec{\sigma})]] = 00101010100000 = f[\phi_{90}(\vec{\sigma})]. \quad (23)$$

Now consider the specific case of SDCA simulating a conventional CA (we follow Majercik [39]). First, because SDCA cannot be expected to preserve the local topology of a simulated CA, it is necessary to define separate *encoding* ( $= e$ ) and *decoding* ( $= d$ ) functions –  $e: \Sigma_{CA} \rightarrow \Sigma_{SDCA}$  transforms the initial configuration of the CA systems to configurations in the SDCA system being used to simulate it (where  $\Sigma_{CA}$  and  $\Sigma_{SDCA}$  are the configurations spaces of CA and SDCA, respectively); and

$d: \Sigma_{SDCA} \rightarrow \Sigma_{CA}$  effectively performs the inverse transformation. Encoding (and decoding) functions are called *structurally defined* if they are recursive and use a finite amount of information to encode (or decode) a given configuration; and are otherwise expected to transform quiescent states to quiescent states. Majercik further assumes that (1)  $e$  has access to the rule table of the conventional CA system being simulated; (2)  $d$  does not have access to the rule tables of either system; and (3)  $e$  and  $d$  must together satisfy the relation:  $e \times d = \text{Identity}(\Sigma_{CA})$ .

Denoting the global transition functions of the CA and SDCA systems by  $\Phi_{CA}$  and  $\Phi_{SDCA}$ , respectively,  $\Phi_{SDCA}$  is said to *simulate*  $\Phi_{CA}$  if there exist  $m \geq 1$ ,  $n \geq 1$  and structurally-defined functions  $e: \Sigma_{CA} \rightarrow \Sigma_{SDCA}$  and  $d: \Sigma_{SDCA} \rightarrow \Sigma_{CA}$ , such that for any configuration  $\vec{\sigma} \in \Sigma_{CA}$  and any  $k \geq 1$ ,

$$\Phi_{CA}^{kn}(\vec{\sigma}) = d \left[ \Phi_{SDCA}^{km} [e(\vec{\sigma})] \right]. \quad (24)$$

If  $m > n$  then  $\Phi_{SDCA}$  simulates  $\Phi_{CA}$  with a *slowdown* factor of  $m/n$ . If  $m < n$  then  $\Phi_{SDCA}$  simulates  $\Phi_{CA}$  with a *speedup* factor of  $n/m$ .

### SDCA Hierarchy of Models

Majercik [39] uses the three generalized models introduced above ( $M_R, M_L$ , and  $M_R$ ) to define a hierarchy of eight SDCA models of computation. At the top of his hierarchy (arranged from top-to-bottom in roughly, but not completely, decreasing order of computational strength; see discussion that follows) are the UL and BL versions of  $M_R$ : SDCA<sup>(8)</sup> and SDCA<sup>(7)</sup>, respectively; followed by SDCA<sup>(6)</sup> = UL version of  $M_L$ ; SDCA<sup>(5)</sup> = BL version of  $M_L$ ; SDCA<sup>(4)</sup> = a *finite labels* version of  $M_L$ ; SDCA<sup>(3)</sup> = UL version of  $M_S$ ; SDCA<sup>(2)</sup> = BL version of  $M_S$ ; and, sitting on the lowest level (computationally speaking), is SDCA<sup>(1)</sup> = SDCA<sub>0</sub>.

A little thought suffices to establish certain relationships among the various classes. Give two classes,  $C_1$  and  $C_2$ , let  $C_1 \leq_s C_2$  denote the fact that if, given any SDCA  $S_1 \in C_1$  there exists an SDCA  $S_2 \in C_2$  that simulates  $S_1$ . Then, for example, since any BL SDCA can be simulated by an unbounded links version of the same system, and a finite links version of  $M_L$  can be simulated by a bounded links version, we know immediately that SDCA<sup>(7)</sup>  $\leq_s$  SDCA<sup>(8)</sup>, SDCA<sup>(4)</sup>  $\leq_s$  SDCA<sup>(5)</sup>  $\leq_s$  SDCA<sup>(6)</sup>, and SDCA<sup>(2)</sup>  $\leq_s$  SDCA<sup>(3)</sup>. Similar reasoning [39] leads to the general relationship:

$$\begin{cases} \text{SDCA}^{(3)} \leq_s \text{SDCA}^{(8)} \leq_s \text{SDCA}^{(6)}, & \text{and} \\ \text{SDCA}^{(2)} \leq_s \text{SDCA}^{(7)} \leq_s \text{SDCA}^{(5)}. \end{cases} \quad (25)$$

Finally, since the unbounded links version of  $M_R$  has all the information necessary to construct the neighborhood partitions used by  $SDCA_0$ , and since  $SDCA^{(8)} \leq_s SDCA^{(6)}$ , we see that  $SDCA_0 \leq_s SDCA^{(6)}$  and  $SDCA_0 \leq_s SDCA^{(8)}$ .

Majercik's two main results, which we state without proof, are:

*Majercik Theorem 1:* Given an arbitrary 1-dimensional conventional CA with radius  $r = 1$ , there exists an unbounded links version of  $M_R (= SDCA^{(8)})$  of the SDCA hierarchy) that can simulate it with a speedup factor of two.

*Majercik Theorem 2:* There exists a 1-dimensional finite links version of  $M_L (= SDCA^{(4)})$  that can simulate an arbitrary  $k$ -state 1-dimensional conventional CA with radius  $r = 1$  with a slowdown factor  $O(k^{2r} \sqrt{2r \log k})$ .

Detailed proofs of these two theorems appear in [39] (where they are called *Theorems 4.4* and *4.5*, respectively). In chapter 5 of his thesis [39], Majercik presents an explicit construction of a CA-universal  $SDCA^{(4)}$  computational model, and compares it to Albert and Culik's [3] construction of a 1-dimensional CA-universal conventional CA that simulates any 1-dimensional,  $k$ -state, radius  $r$  CA with an  $O(k^{8r})$  slowdown. Although Majercik's CA-universal SDCA uses more states than Albert and Culik's universal CA, it is also markedly faster.

The reason why the SDCA is faster is at least intuitively clear. An SDCA's dynamic links effectively endow an otherwise conventional CA with a random access memory. Since SDCA can establish links between any two sites a distance  $d$  apart in  $O(\log d)$  time, any site potentially has access to the state of any other site. While it may be argued that sites in conventional CA can also access the states of other cells, they *cannot do so permanently*. Once information is accessed once and used, the connection is lost, and must subsequently be re-established. Moreover, the links in SDCA can potentially connect sites that are arbitrarily far apart; so that, once a small number of links are dynamically created, they continue to provide long-range communication channels throughout the network. Since the propagation of information in a conventional CA is necessarily limited in being able to flow one site at a time, the overall computational speed is obviously limited.

However, it is worth pointing out that while the computational strength of Majercik's CA-universal SDCA model undoubtedly derives from its ability to forge long-range communication links, the results as quoted

from [39] do not tap into what is potentially SDCA's greatest strength; namely, the ability to *adaptively create links, even as a given computation unfolds*. In Majercik's model, the links are dynamically coupled to an actual computation only insofar as they are initially fixed as a function of the initial state. While the local structure certainly *evolves* (as it does in all SDCA systems, as the computation itself unfolds), it does so purely as a consequence of the SDCA rules, and not adaptively to the evolution.

Majercik concludes his thesis by speculating on how an *adaptive* variant of his CA-universal SDCA may be used to explore certain aspects of evolutionary learning. (Working from a different set of assumptions, Halpern [23,24] applies evolutionary programming techniques to  $SDCA_0$  to explore what happens when the structure is allowed to play an *explicit* dynamic role in the computation; see next section.) The question of whether there exist SDCA-universal SDCA models – that are able to simulate certain classes of the SDCA hierarchy, for example – remains open.

## SDCA & Genetic Algorithms

*Genetic algorithms* (abbreviated, GA) are a class of heuristic search algorithms and computational models of adaptation and evolution based on natural selection. In nature, the search for beneficial adaptations to a continually changing environment (i. e., evolution) is fostered by the cumulative evolutionary knowledge that each species possesses of its forebears. This knowledge, which is encoded in the chromosomes of each member of a species, is passed on from one generation to the next by a mating process in which the chromosomes of “parents” produce “offspring” chromosomes. A comprehensive review of GA is given by Mitchell [45].

While GAs may be effectively used to search for “interesting” topological structures (but for which the structures themselves do not play any dynamic role; see, for example, Lehmann [37]), Halpern [23] is the first to explore a novel hybrid algorithm between GA and SDCA, in which SDCA rules are used to *evolve* a GA. Weinert et al. [72] explore a related “structurally dynamic” GA model, in which links between adjacent individuals of a population are dynamically chosen according to deterministic or probabilistic rules. In this section, we follow Halpern [23,24].

Formally, GAs are defined by (1) an ensemble of “candidate solution” vectors,  $\{\vec{s}_i\}: \vec{s}_i \in M_P \subseteq R^n$ , where  $M$  is the set of all possible solutions to a given “problem”  $\mathcal{P}$  (the  $\vec{s}_i$  are usually, but not always, defined as a string of binary numbers [45]), and (2) a “fitness function”,  $f(\vec{s})$ , that represents how well a given  $\vec{s}$  “solves”  $\mathcal{P}$ . The goal of the GA is to find the global optimal solution,

$\vec{s}^*$  such that (from the point of view of *maximizing* fitness,  $f(\vec{s}) \leq f(\vec{s}^*) \equiv f^*, \forall \vec{s} \in M$ ). Optimization proceeds through the combined processes of *selection*, *breeding*, *mutation*, and *crossover replacement* [45]; to which – in the hybrid SDCA $\leftrightarrow$ GA algorithm, Halpern adds the new feature of *self-selective neighborhood structure*.

It should be immediately noted that this is not an ad-hoc addition. Muhlenbein [47] points out that if each generation of a GA searches over the entire possible solution space, the algorithm may – depending on the fitness function – converge prematurely to a sub-optimal solution. To reduce the likelihood of this happening, Muhlenbein introduces a spatial population structure; restricting fitness and mating to neighborhoods called *demes*. Demes are geographically separate subpopulations in which candidate solutions evolve along disparate trajectories; though occasional mixing still occurs through the process of migration.

In Halpern's variant [23], an otherwise conventional GA is placed within the structure of SDCA<sub>0</sub> (i. e., the basic model defined by Eqs. (11)–(14)). Heuristically, this allows each candidate solution to “choose a community” with which to mate, during each generation. The choice of neighborhoods thus becomes an integral component of the GA, and is determined dynamically by the evolving solutions.

Halpern's algorithm proceeds as follows [24]: (*Step 1*) an initially random lattice (defined by adjacency matrix  $l_{ij}^{(t=0)}$ ) is seeded with single-chromosome candidate solutions of fixed length, one per site; (*Step 2*) a fitness function,  $f_i = \sum_{j=1}^N \delta_{ij}$ , is defined to assign a numerical measure of “optimality” to each site ( $N$  is the number of sites, and  $\delta_{ij}$  is the value – equal to 0 or 1 – of the  $j$ th gene of the  $i$ th chromosome; (*Step 3*) each site  $i$  ranks each of its nearest and *next*-nearest neighbors according to  $f_j$ ; (*Step 4*) each site disconnects with a fraction,  $f_D$ , of its least-fit neighbors, and connects with a fraction,  $f_C$ , of its fittest next-nearest neighbors; (*Step 5*) each site randomly mates with one of its nearest neighbors (i. e., the usual processes of mutation and crossover operations are applied [45]); (*Step 6*) the least fit members of the population are replaced by the offsprings from *Step 5*; and (*Step 7*) loop through steps 5–7, until some suitable “optimality” threshold (or some other convergence criterion) is satisfied.

Halpern [23,24] reports a wide range of resulting behaviors, collectively suggesting a clear relationship between the parameters defining the GA optimization and lattice connectivity. Of particular interest are the dynamic conditions for which the fitness-based creation and deletion of links increases the rate of growth of overall fitness. The fastest convergence occurs when lattice connectivity

first increases, then decreases, then eventually levels off. In the first stage, the fittest possible communities are first established; in the second stage, connections with poorer candidate solution are deleted; finally, in the third stage, the system essentially “fine-tunes” its optimal solutions. Halpern [24] finds two different evolutionary paths toward high connectivity: (1) monotonic growth over time (for low mutation rates,  $p_\mu$ ), and (2) a *phase transition* between low and high degrees of connectivity (for some  $p_\mu^*$ ). Using SDCA $\leftrightarrow$ GA hybrid model parameters  $N = 100$  and  $f_D = f_C = 0.1$ ,  $p_\mu^* \approx 0.05$ , for which Halpern [24] finds a sharp increase in the number of links per site between generations 350 and 450.

Despite the novelty of the approach, and the promising link between optimization rates and dynamic structure established in [23], concrete applications of the algorithm – except for Weinert et al. [72] work on a related hybrid GA algorithm – have yet to be developed. One suggestion, from Halpern [24], is to use the SDCA $\leftrightarrow$ GA hybrid model for finding “optimal” connectivity patterns in parallel computers. The search algorithm may be used to directly model how component processors are connected, and decide to keep or sever existing links, or establish new ones, adaptively as a function of local fitness criteria.

### Generalized SDCA Models

Despite SDCA being obviously more “complex” than conventional CA (and certainly more complex to formally define, if only because one must specify *both*  $\sigma$  and  $\ell$  rules), the SDCA model nonetheless has more in common with *elementary* CA than with any of its brethren's more “complicated” variants. By “elementary” CA we mean the simplest one-dimensional CA with  $\sigma \in \{0, 1\}$  and local neighborhoods consisting only of left and right (i. e., *nearest*) neighbors. Just as there are many generalizations of elementary CA – for example, increasing the state space to include  $\sigma$ 's that take on one of  $N$  values, larger-sized neighborhoods, and memory, among many other possibilities – so too there are natural extensions of basic SDCA. In this section we discuss three generalizations: (1) rules that are *reversible in time*, (2) rules that retain a *memory of past states*, and (3) *probabilistic* rules.

### Reversible SDCA

The first generalization of the basic SDCA model, explored extensively by Alonso-Sanz [7], is to apply the *Fredkin reversible-rule construction* to  $\ell$  rules to render them reversible in time. Consider a conventional CA system that is first-order in time,  $\sigma_i^{t+1} = \phi[\sigma_j^t \in \mathcal{N}_i]$ , where  $\mathcal{N}_i$  is the

neighborhood around site  $i$  and, generally,  $\sigma_i \in \mathbb{Z}_k$ . The *Fredkin* construction converts this system into an explicitly invertible one that is *second-order* in time by subtracting the value of the center site at time  $t - 1$ :

$$\sigma_i^{t+1} = \phi \left[ \sigma_j^t \in \mathcal{N}_i \right] \ominus_k \sigma_i^{t-1}, \quad (26)$$

where ' $\ominus_k$ ' is subtraction *modulo-k*. Since Eq. (26) can be trivially solved for  $\sigma_i^{t-1}$  ( $= \phi[\sigma_j^t \in \mathcal{N}_i] \ominus_k \sigma_i^{t+1}$ ), we see that any pair of consecutive configurations uniquely specifies the backwards trajectory of the system. Moreover, this is true for arbitrary (and, in particular, irreversible) functions  $\phi$ .

Now, exactly the same procedure may be applied to link functions:

$$\begin{cases} l_{ij}^{t+1} = \psi \left( \{\sigma_k^t\}, \{l_{ij}^t\} \right) \ominus_2 l_{ij}^{t-1}, \\ l_{ij}^{t+1} = \omega \left( \{\sigma_k^t\}, \{l_{ij}^t\} \right) \ominus_2 l_{ij}^{t-1}, \end{cases} \quad (27)$$

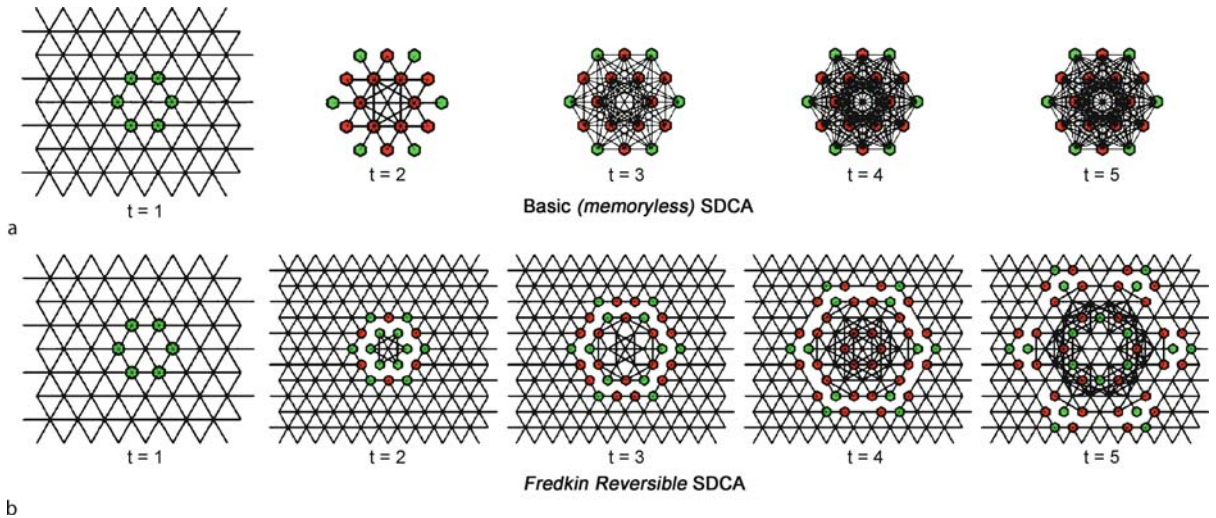
where  $\ominus_2$  is subtraction *modulo-2* (since links are obviously binary valued).

Following Alonso-Sanz [7,8], we consider these two specific SDCA link rules (which will also be used in a later example):

$$\begin{cases} \psi \left( \sigma_i^t, \sigma_j^t, l_{ij}^t \right) = 0 \text{ iff } l_{ij}^t = 1 \text{ and } \sigma_i^t + \sigma_j^t = 0, \\ \omega \left( \sigma_i^t, \sigma_j^t, l_{ij}^t \right) = 1 \text{ iff } l_{ij}^t = 0, \sigma_i^t > 0, \sigma_j^t > 0, \text{ and } D_{ij} = 2. \end{cases} \quad (28)$$

Figure 10 compares the evolution of the *Fredkin reversible* version of these rules to their memoryless counterpart. Both evolutions start on a two dimensional hexagonal lattice, and values evolve according to the three-state (i.e.,  $\sigma \in \{0, 1, 2\}$ ), next-nearest neighborhood **T beehive** rule. The *beehive* rule is defined explicitly by assigning one of three values (0, 1, or 2), to each possible 3-tuple,  $(N_0, N_1, N_2)$ , that gives the number of local sites with  $N_0$  0s,  $N_1$  1s, and  $N_2$  2s [7]:  $(0, 0, 6) \rightarrow 0$ ,  $(0, 1, 5) \rightarrow 1$ ,  $(0, 2, 4) \rightarrow 2$ ,  $(0, 3, 3) \rightarrow 1$ ,  $(0, 4, 2) \rightarrow 2$ ,  $(0, 5, 1) \rightarrow 0$ ,  $(0, 6, 0) \rightarrow 0$ ,  $(1, 0, 5) \rightarrow 0$ ,  $(1, 1, 4) \rightarrow 2$ ,  $(1, 2, 3) \rightarrow 2$ ,  $(1, 3, 2) \rightarrow 2$ ,  $(1, 4, 1) \rightarrow 1$ ,  $(1, 5, 0) \rightarrow 1$ ,  $(2, 0, 4) \rightarrow 0$ ,  $(2, 1, 3) \rightarrow 0$ ,  $(2, 2, 2) \rightarrow 2$ ,  $(2, 3, 1) \rightarrow 2$ ,  $(2, 4, 0) \rightarrow 0$ ,  $(3, 0, 3) \rightarrow 0$ ,  $(3, 1, 2) \rightarrow 2$ ,  $(3, 2, 1) \rightarrow 2$ ,  $(3, 3, 0) \rightarrow 0$ ,  $(4, 0, 2) \rightarrow 0$ ,  $(4, 1, 1) \rightarrow 0$ ,  $(4, 2, 0) \rightarrow 2$ ,  $(5, 0, 1) \rightarrow 2$ ,  $(5, 1, 0) \rightarrow 0$ ,  $(6, 0, 0) \rightarrow 0$ .

The top row of Fig. 10 shows the first four steps ( $t = 1, 2, 3$ , and 4) in the memoryless evolution of the initial "ring" of sites that appears at  $t = 1$ . The link rules used for this run are those defined in Eq. (28). Since the decoupler removes links between pairs of sites whose values are equal to zero, most of the lattice disappears after a single time step, and both value and link activity is confined to a small region. After two more steps of changes, the system quickly attains a fixed point:  $\{\sigma^t, \ell_{ij}^t\} = \{\sigma^{t=4}, \ell_{ij}^{t=4}\}$  for all  $t \geq 5$ . While the frequency of states is not constrained to total six for a dynamic lattice, the *beehive* rule is unchanged; if the sum of frequencies at a given site exceeds six, the site value remains the same.



Structurally Dynamic Cellular Automata, Figure 10

Comparison between first few time steps of **a** a memoryless SDCA, evolving according to link rules defined in Eq. (28), and **b** the *Fredkin reversible* versions of these rules (obtained by applying Eq. (27) to Eq. (28)). In both cases,  $\sigma$ 's evolve according to the *beehive* rule defined in the text. (Reproduced with permission from [7])



The bottom row shows the evolution of the *Fredkin reversible* versions of the rules defined in Eq. (28) (to simplify the visualization, links along the border sites are not shown). In contrast to the basic SDCA version, the initial lattice in this case *does not* decay. Since, according to Eq. (27) (which assumes that  $\ell_{ij}^{t=0} = \ell_{ij}^{t=1}$ ), the initial hexagonal lattice is subtracted from the evolved structure at  $t = 1(\text{modulo-2})$ , the original graph is effectively restored, and the outlying regions appear undisturbed.

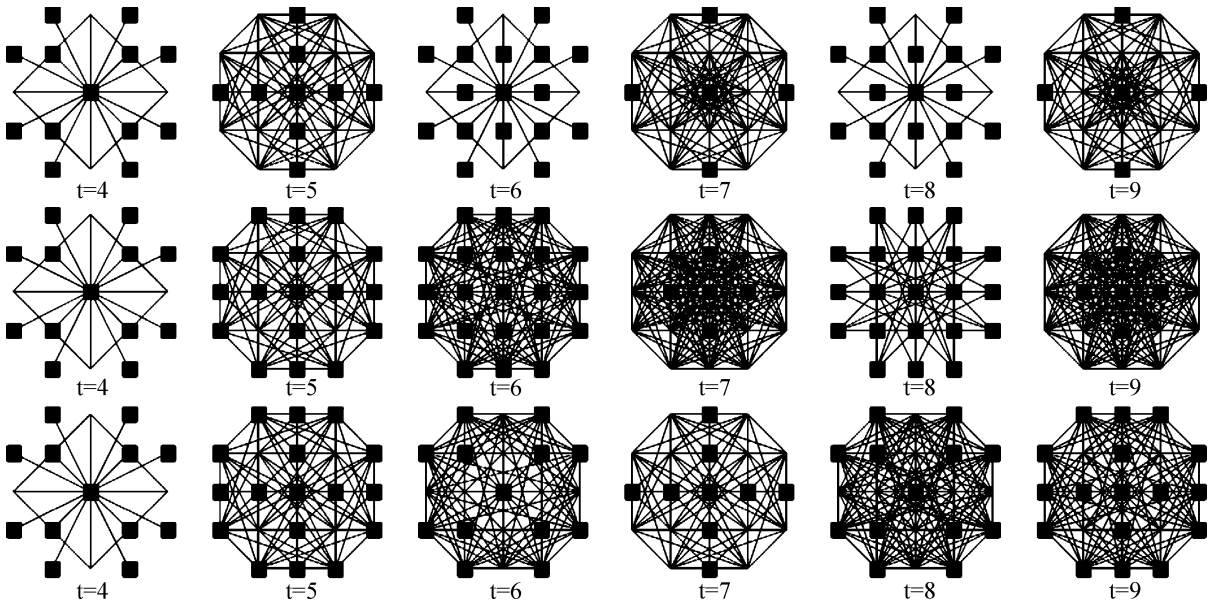
### SDCA with Memory

A second generalization to the basic SDCA model, introduced and studied by Alonso-Sanz and Martín [6,7,8], is to endow both  $\sigma$ -rules and  $\ell$ -rules with *memory*. The rules for conventional memoryless CA and SDCA, depend only on neighborhood configurations that appear on the immediately preceding time step. Therefore, rules may be said to possess a “memory” of depth  $m$  if they depend explicitly on values (in the case of CA), or on both values and link states (in the case of SDCA), that existed on  $m$  previous time steps. We note, in passing, that since the *Fredkin construction* couples states at times  $t + 1$ ,  $t$  and  $t - 1$ , reversibility may be considered a specific form of memory that extends backwards a single step.

Of course, there is no unique prescription for introducing a dependency on past values; and a variety of alternative memory mechanisms have been proposed in the literature (for example, see page 43 in [32] and page 118 in [74]). We focus our discussion on the approach proposed by Alonso-Sanz (14), and for the moment confine our attention to value rules,  $\phi: \sigma \rightarrow \sigma'$ . Alonso-Sanz’s approach is to preserve the form of the transition rule, but have it act on an effective site value that is a weighted function of its  $m$  prior values.

This is done by introducing a memory-endowed value rule,  $\phi_m$ , that – in contrast to its memoryless version,  $\phi$  – is not, in general, a function of a given site’s current value,  $\sigma_i$ , *alone*, but is instead a function of the transformed value,  $s = \mathcal{M}_\phi(\sigma; m, \alpha)$ , obtained from  $\sigma_i$ ’s past  $m$  values:  $\phi_m: s \rightarrow \sigma'$ , where  $0 \leq \alpha \leq 1$  is a numerical *memory factor*. The value transforming memory function,  $\mathcal{M}$ , assumes the following specific form (to avoid confusion, note that in Eqs. (29) and (30),  $\sigma_i^x$  means *the value of  $\sigma_i$  at time  $t = x$* , and  $\alpha^x$  means *the numerical quantity  $\alpha$  raised to the power  $x$* ):

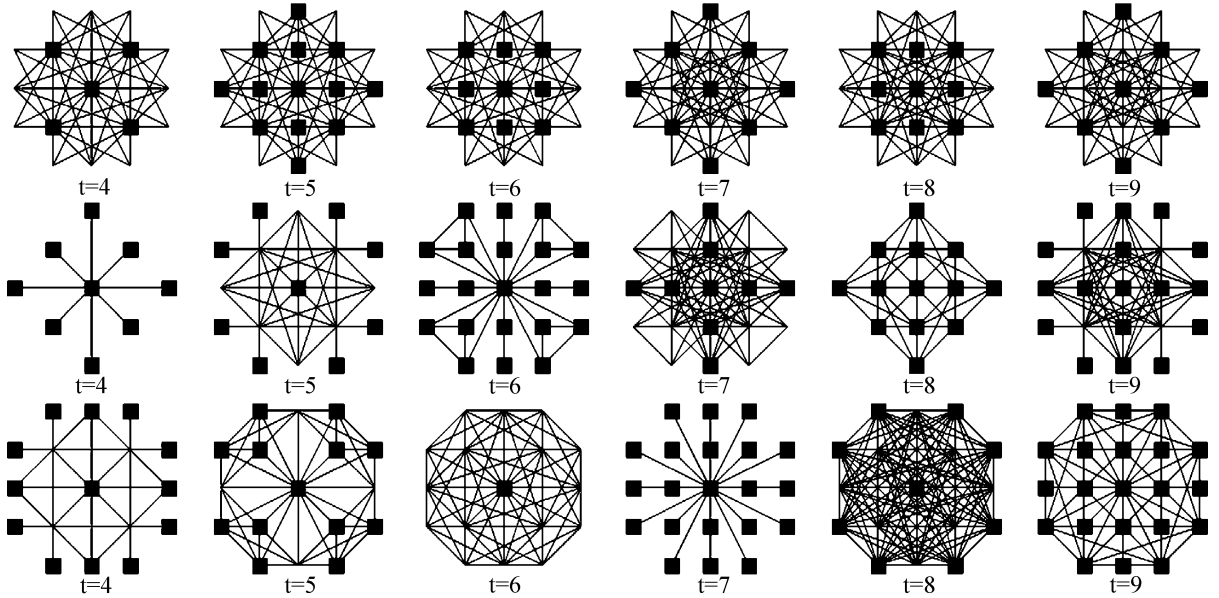
$$s_i^t = \mathcal{M}_\phi(\sigma_i^t; m, \alpha) = \begin{cases} 1 & \text{if } (\hat{\sigma}_i^t)_m > 1/2, \\ \sigma_i^t & \text{if } (\hat{\sigma}_i^t)_m = 1/2, \\ 0 & \text{if } (\hat{\sigma}_i^t)_m < 1/2, \end{cases} \quad (29)$$



Structurally Dynamic Cellular Automata, Figure 11

Sample runs of a SDCA with memory for memory weighting  $\alpha = 0.6$ . The SDCA is initialized as a Euclidean four-neighbor lattice, and evolves according to the *parity T*  $\sigma$ -rule and the two  $\ell$  rules defined in Eq. (28). The first row of evolving patterns applies memory only to *values*; the second row applies memory only to *links*, and the third row shows the evolution when memory is applied to both. (Reproduced by permission from [8])





Structurally Dynamic Cellular Automata, Figure 12

Sample runs of the same SDCA shown in Fig. 11, but with memory weighting  $\alpha = 1.0$ . (Reproduced by permission from [8])

where

$$(\hat{\sigma}_i^t)_m = \frac{\sigma_i^t + \sum_{\Delta t=1}^m \alpha^{\Delta t} \cdot \sigma_i^{t-\Delta t}}{1 + \sum_{\Delta t=1}^m \alpha^{\Delta t}}. \quad (30)$$

At any given time,  $t$ , the depth  $m$  can never exceed  $t - 1$ . Our discussion follows Alonso-Sanz [8], and sets  $m(t) \equiv t - 1$  for all  $t$ ; i.e., we assume that  $\mathcal{M}_\phi(\sigma; m, \alpha)$  yields a weighted mean value of *all* the previous values of a given site. In practice, memory becomes active only after a certain number of initialization steps, here taken to be three; with seeded values  $s_i^1 = \sigma_i^1$  and  $s_i^2 = \sigma_i^2$ .

Memory can be added to link rules in a similar manner. The form of the link rules ( $\psi$  and  $\omega$ ) remains the same, but rather than acting on a graph that is defined by its adjacency matrix,  $\ell_{ij}^t$ ,  $\psi$  and  $\omega$  instead act on the memory-transformed values,  $L = \mathcal{M}_{(\psi, \omega)}(\ell; m, \alpha)$ :

$$L_{ij}^t = M_{(\psi, \omega)}(\ell_{ij}^t; m, \alpha) = \begin{cases} 1 & \text{if } (\hat{\ell}_{ij}^t)_m > 1/2, \\ \ell_{ij}^t & \text{if } (\hat{\ell}_{ij}^t)_m = 1/2, \\ 0 & \text{if } (\hat{\ell}_{ij}^t)_m < 1/2, \end{cases} \quad (31)$$

where

$$(\hat{\ell}_{ij}^t)_m = \frac{\ell_{ij}^t + \sum_{\Delta t=1}^m \alpha^{\Delta t} \cdot \ell_{ij}^{t-\Delta t}}{1 + \sum_{\Delta t=1}^m \alpha^{\Delta t}}. \quad (32)$$

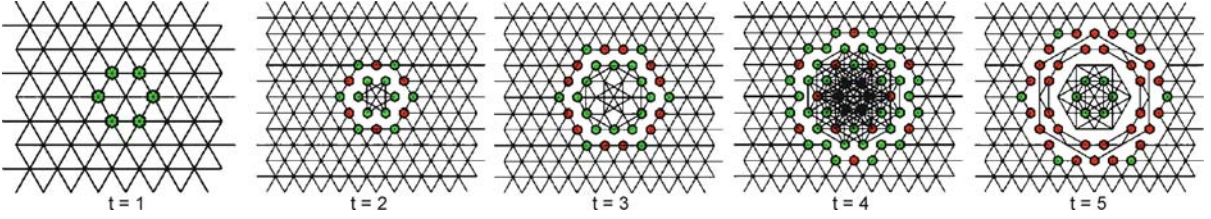
As for memory-endowed  $\sigma$ -rules, the memory for link rules is activated only on the third iteration step, and the system is initialized by setting  $L_i^1 = \sigma_i^1$  and  $L_i^2 = \sigma_i^2$ .

Figures 11 and 12 show the effects of applying partial memory weighting ( $\alpha = 0.6$ ) and full memory ( $\alpha = 0.6$ ), respectively, to a SDCA that starts with a Euclidean four-neighbor lattice, and evolves according to the *parity T*  $\sigma$ -rule (that assigns a value *zero* to a site if the sum of the values in its neighborhood is *even*, and assigns the value *one* if the sum is *odd*) and the  $\ell$  rules defined above in Eq. (28). The first row of evolving patterns (for each  $\alpha$ ) applies memory only to *values*; the second applies memory only to *links*, and the third applies memory to both. Figure 13 shows the reversible *beehive* SDCA shown in Fig. 10, but with full memory ( $\alpha = 1.0$ ).

### Probabilistic SDCA

Another natural extension of the basic SDCA model is to replace the set of explicit  $\sigma$ - and/or  $\ell$ -rules with probabilities. In this way one can study the evolution of a system that undergoes random but  $\sigma$ -dependent lattice changes. For example, this may be useful for studying genetic networks in which new links are forged (with a given probability) only if both genes are active, and existing connections are broken if both sites are inactive.

Following Halpern and Caltagirone [22], consider the *parity T*  $\sigma$ -rule and the following probabilistic versions of



Structurally Dynamic Cellular Automata, Figure 13

Sample runs of a reversible *beehive* SDCA with full memory ( $\alpha = 1.0$ ); compare to Fig. 10. (Reproduced by permission from [8])

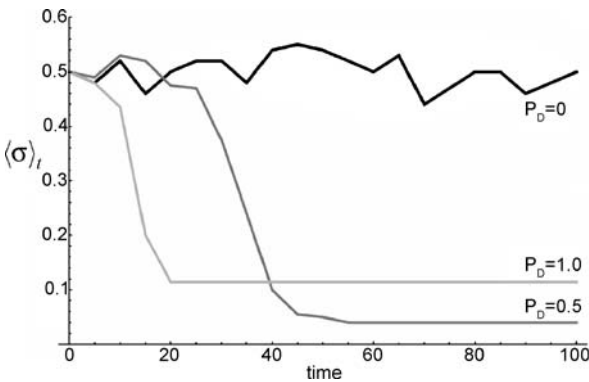
decoupler ( $\psi_p$ ) and coupler rules ( $\omega_p$ ):

$$\begin{aligned}
 (\text{decoupler}) : & \begin{cases} l_{ij}^{t+1} = \psi_p(l_{ij}^t; \sigma_i^t, \sigma_j^t, p_D), \\ \psi_p \equiv 1 - \delta(\sigma_i^t + \sigma_j^t, 0) \cdot \delta(p_D > r), \end{cases} \\
 (\text{coupler}) : & \begin{cases} l_{ij}^{t+1} = \omega_p(l_{ij}^t; \sigma_i^t, \sigma_j^t, p_D), \\ \omega_p \equiv \delta(D_{ij}, 2) \cdot \delta(\sigma_i^t + \sigma_j^t, 2) \cdot \delta(p_C > r), \end{cases}
 \end{aligned} \quad (33)$$

where  $p_D$  and  $p_C$  are the *decoupler* and *coupler* probabilities, respectively, and  $r$  is a random number between 0 and 1.

Thus,  $\psi_p$  unlinks two previously linked sites with probability  $p_D$  if and only if the sum of their site values is zero; and  $\omega_p$  links two previously unlinked sites with probability  $p_C$  if and only if they are next-nearest neighbors and the sum of their site values is two.

Figure 14 shows time series plots of  $\langle \sigma \rangle$  as a function of time for three different cases: (1)  $p_D = 0$  (no decoupling at all), (2)  $p_D = 1/2$ , and (3)  $p_D = 1$  (decoupler rule applied 100% of the time (consistent with *non-probabilistic* SDCA rules). We see that changing  $p_D$  induces qualitatively different  $\sigma$  behavior, that ranges from small fluctu-



Structurally Dynamic Cellular Automata, Figure 14

Time series of average  $\sigma$  value,  $\langle \sigma \rangle_t$ , for the Halpern-Caltagirone rules (defined in Eq. (33)) and for three values of *decoupler* probability:  $p_D = 0$ ,  $p_D = 1/2$ , and  $p_D = 1$ . (Reproduced with permission from [22])

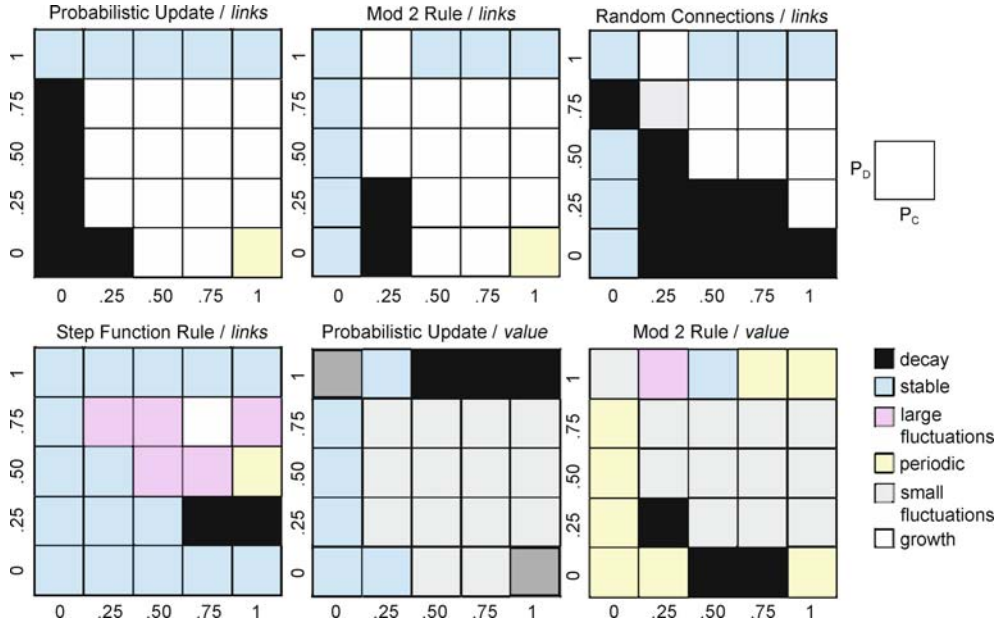
ations around  $\langle \sigma \rangle \sim 0.5$  (for  $p_D = 0$ ), to decay to small static values ( $\langle \sigma \rangle = 0.05$  for  $p_D = 1/2$ , and  $\langle \sigma \rangle = 0.12$  for  $p_D = 1$ ).

Halpern and Caltagirone [22] have studied a wide range of probabilistic SDCA, using random initial  $\sigma$  configurations, *step-function*, *parity*, and Conway's *life*  $\sigma$ -rules, Cartesian and random initial lattice structures, and various probabilities  $0 \leq p_D \leq 1$  and  $0 \leq p_C \leq 1$ . Some of their results are reproduced (with permission) in the *behavioral phase* plots shown in Fig. 15. (The *step-function* rule is defined by  $\sigma_i^{t+1} = 0$  if and only if  $\sum_j \ell_{ij}^t \sigma_j^t > 2$  and  $\sigma_i^{t+1} = 1$  if and only if  $\sum_j \ell_{ij}^t \sigma_j^t \leq 2$ ; Conway's *life* rule assigns  $\sigma_i^{t+1} = 1$  to a site if and only if  $\sigma_i^{(t)} = 0$  and the sum of values in its neighborhood at time  $t$  is equal to 3 or  $\sigma_i^{(t)} = 1$  and the sum of values is equal to 2 or 3; otherwise  $\sigma_i^{t+1} = 0$ .)

Figure 15 shows a wide range of possible behaviors. Consider, for example, the number of *links per site* for the case where the lattice is updated with probabilistic  $\ell$ -rules and the  $\sigma$ 's are all random (shown at the top left of the figure). Four distinct classes of behavior appear, with *growth* dominant for most values of  $p_D$  and  $p_C$ .

Pure decoupling (or pure coupling) leads to complete decay (or growth to a stable state); a mixed state of coupling/decoupling generally yields slow growth. Periodic behavior occurs only for  $p_D \sim p_C \sim 1$ . Compare this behavior with the cases where the  $\sigma$ -rule is either the *parity* value rule (shown in the middle of the top row of Fig. 15) or the *step-function* rule (shown at left bottom of the figure). While the *parity* rule also displays four similar phases (growth to stability, decay to stability, incomplete growth, and incomplete decay), decaying structures eventually reach a stable (not null) final state. The *step-function* rule shows an even greater variety of possible behaviors, and appears more sensitive to small changes in link probabilities.

The probabilistic SDCA system discussed in this section adds a stochastic element *specifically to SDCA*. Of course, there are other ways of injecting stochasticity into a CA with dynamic topology. For example, Makowiec [40] combines the deterministic evolution of a conventional



**Structurally Dynamic Cellular Automata, Figure 15**

Behavioral phase plots summarizing the long term evolution for several different  $\sigma$  and  $\ell$ -rules defined in Eq. (33). The x and y axes for each plot depict values ( $\in \{0, .25, .5, .75, 1\}$ ) of  $P_C$  and  $P_D$ , respectively. There are six classes of behavior: *growth*, *decay*, *stability*, *large* and *small fluctuations* (around a stable lattice), and *periodicity*. The initial graph is a Cartesian four-neighbor lattice in each case except for the *top-right* plot (labeled *Random connections/links*) for which the initial graph is random. (Reproduced with permission from [22])

CA with an asynchronous stochastic evolution of its underlying lattice (patterned after the Barabasi–Albert [9] model of degree distributions in small-world networks), to explore the influence of dynamic topology on the zero-temperature limit of ferromagnetic transitions.

### Random Dynamics Approximation

For cases in which the structure and value configurations are both sufficiently random and uncorrelated, a *random dynamics approximation* (abbreviated, RDA) may suffice to qualitatively predict how the system will tend to evolve under a specific rule set; for example, to predict whether a given rule is more (or less) likely to yield unbounded growth, to eventually settle into a low periodic state, or to simply decay. The idea is to approximate the real SDCA as a *mean-field*; that is, assume all local value and structural correlations are close to zero (and can thus be ignored), and replace all specific site values and local link geometries with average, or effective, values.

More precisely, assuming that (1) the probability  $p_n^{(\sigma_i)}$  of a site ' $i$ ' having value  $\sigma = 1$  at time  $t = n$  is the same for all sites – so that  $p_n^{(\sigma_i)} = p_n^{(\sigma)}$  for all  $i$  – and (2) that the probability  $p_n^{(l_{ij})}$  of two sites ' $i$ ' and ' $j$ ' being linked at  $t = n$  is the same for all pairs of sites – so that  $p_n^{(l_{ij})} = p_n^{(l)}$  for all  $i$  and  $j$  – the RDA evolution equations may be written

formally as follows:

$$\begin{cases} p_{n+1}^{(\sigma)} = F_{\text{RDA}}[p_n^{(\sigma)}, p_n^{(l)}; \Gamma_{\text{SDCA}}], \\ p_{n+1}^{(l)} = G_{\text{RDA}}[p_n^{(\sigma)}, p_n^{(l)}; \Gamma_{\text{SDCA}}], \end{cases} \quad (34)$$

where SDCA's rule  $\Gamma_{\text{SDCA}}$  (defined in Eq. (14)) is included, formally, to remind us that the functional forms assumed by  $F_{\text{RDA}}$  and  $G_{\text{RDA}}$  will be different for different  $\Gamma_{\text{SDCA}}$ s.

The first function,  $F_{\text{RDA}}$ , is the easier of the two to calculate. For any given site with degree  $d$  we simply count the total number of ways to distribute the local  $\sigma$ -values among the  $d$  possible neighboring sites to obtain the desired sums that define a given rule. In this way we find the average expected  $\sigma$  density at  $t = n + 1$ , assuming all sites in the lattice have the same degree  $d$  at time  $t = n$ :

$$\begin{aligned} p_{n+1}^{(\sigma)}(d, p_n^{(\sigma)}) &= \begin{cases} \sum_{\{\alpha\}} \binom{d+1}{\alpha} [p_n^{(\sigma)}]^\alpha (1 - p_n^{(\sigma)})^{d+1-\alpha} & \Leftrightarrow \mathbf{T} \\ \sum_{\{\alpha_0\}} \binom{d}{\alpha_0} [p_n^{(\sigma)}]^{\alpha_0+1} (1 - p_n^{(\sigma)})^{d-\alpha_0} \\ \quad + \sum_{\{\alpha_1\}} \binom{d+1}{\alpha_1} [p_n^{(\sigma)}]^{\alpha_1} (1 - p_n^{(\sigma)})^{d+1-\alpha_1} & \Leftrightarrow \mathbf{OT} \end{cases} \quad (35) \end{aligned}$$

We then get  $F_{\text{RDA}} \rightarrow p_{n+1}^{(\sigma)} = \sum_d P(d; p_n^{(l)}) \cdot p_n^{(\sigma)}(d, p_n^{(\sigma)})$  as an average over all possible degrees, where  $P(d; p_n^{(l)})$  is the probability that any site has exactly  $d$  neighbors. Since this means that, out of a total of  $N - 1$  possible neighbors, a given site must have exactly  $d$  links, and not be connected to any of the remaining  $(N - 1 - d)$  sites, we have by inspection:

$$P(d; p_n^{(l)}) = \binom{N-1}{d} [p_n^{(l)}]^d (1 - p_n^{(l)})^{N-1-d}. \quad (36)$$

To calculate the second function in Eq. (34) ( $= G_{\text{RDA}}$ ), we first define the local transition functions

$$\begin{cases} p_n^a(d_1, d_2, \lambda) \\ = \text{Prob}(l = 1 \rightarrow l' = 0 \mid d_i = d_1, d_j = d_2, |A_{ij}| = \lambda), \\ p_n^b(d_1, d_2, \lambda) \\ = \text{Prob}(D = 2 \rightarrow l' = 1 \mid d_i = d_1, d_j = d_2, |A_{ij}| = \lambda), \end{cases} \quad (37)$$

which give the probabilities that any two sites  $-i$  and  $j-$  will be *disconnected* ( $p_n^a$ ) or *connected* ( $p_n^b$ ) if they have prescribed degrees  $d_i = d_1$  and  $d_j = d_2$ , and are each linked to the *same*  $\lambda$  sites in the shared neighbor set,  $A_{ij}$  (see Fig. 1). In the case of type-**T**  $\sigma$ - and  $\ell$ -rules,  $p_n^a$  and  $p_n^b$  are given explicitly by (**OT** versions of  $\sigma$ - and  $\ell$ -rules, and **RT** versions of  $\ell$ -rules are defined by similar, but slightly more complicated, expressions):

$$\begin{cases} p_n^a(d_1, d_2, \lambda) = \sum_k \binom{d_1 + d_2 - \lambda}{\beta_k} [p_n^{(\sigma)}]^{\beta_k} \\ \quad \cdot (1 - p_n^{(\sigma)})^{d_1 + d_2 - \lambda - \beta_k}, \\ p_n^b(d_1, d_2, \lambda) = \sum_k \binom{d_1 + d_2 + 2 - \lambda}{\epsilon_k} [p_n^{(\sigma)}]^{\epsilon_k} \\ \quad \cdot (1 - p_n^{(\sigma)})^{d_1 + d_2 + 2 - \lambda - \epsilon_k}, \end{cases} \quad (38)$$

where  $\beta_k$  and  $\epsilon_k$  refer to the sums that appear in Eqs. (11) and (12). The total probability that any two sites will be disconnected ( $l = 1 \rightarrow l' = 0$ ) or connected ( $D = 2 \rightarrow l' = 1$ ) –  $P_n^a$  and  $P_n^b$ , respectively – may then be obtained by summing over all possible local topologies:

$$\begin{cases} P_n^a \equiv \langle \text{Prob}(l = 1 \rightarrow l' = 0) \rangle \\ = \sum_{d_1} \sum_{d_2} \sum_{\lambda} P_1(d_1, d_2, \lambda) \cdot p_n^a(d_1, d_2, \lambda), \\ P_n^b \equiv \langle \text{Prob}(D = 2 \rightarrow l' = 1) \rangle \\ = \sum_{d_1} \sum_{d_2} \sum_{\lambda} P_2(d_1, d_2, \lambda) \cdot p_n^b(d_1, d_2, \lambda), \end{cases} \quad (39)$$

where

$$\begin{cases} P_1(d_1, d_2, \lambda) = \text{Prob}(\text{sites } i, j \mid l_{ij} = 1 \text{ have } d_i = d_1, \\ \quad d_j = d_2, |A_{ij}| = \lambda), \\ P_2(d_1, d_2, \lambda) = \text{Prob}(\text{sites } i, j \mid l_{ij} = 0 \text{ have } d_i = d_1, \\ \quad d_j = d_2, |A_{ij}| = \lambda). \end{cases} \quad (40)$$

To find  $P_1$  we need to count, from among the remaining  $N - 2$  sites, the number of ways of selecting disjoint sets  $S_1$ , containing  $d_1 - 1 - \lambda$  sites linked only to  $i$ ;  $S_2$ , consisting of  $d_2 - 1 - \lambda$  sites connected only to  $j$ ; and  $S_3$ , with  $\lambda$  sites linked to *both*  $i$  and  $j$ . But this is simply a multinomial coefficient, so we can write:

$$P_1(d_1, d_2, \lambda) = \frac{(N-2)_{d_1+d_2-\lambda-2}}{(d_1-1-\lambda)!(d_2-1-\lambda)!\lambda!} \cdot [p^{(l)}]^{d_1+d_2-2} (1-p^{(l)})^{2(N-1)-d_1-d_2}, \quad (41)$$

where  $(n)_k \equiv n(n-1)\cdots(n-k+1)$ . Similarly, for  $P_2$ , we need to count the number of ways of choosing  $d_1 - \lambda$  sites from  $i$ ,  $d_2 - \lambda$  sites from  $j$ , and  $\lambda$  sites from both:

$$P_2(d_1, d_2, \lambda) = \frac{(N-3)_{d_1+d_2-\lambda-2}}{(d_1-\lambda)!(d_2-\lambda)!\lambda!} \cdot [p^{(l)}]^{d_1+d_2} (1-p^{(l)})^{2(N+\lambda-3)-d_1-d_2}. \quad (42)$$

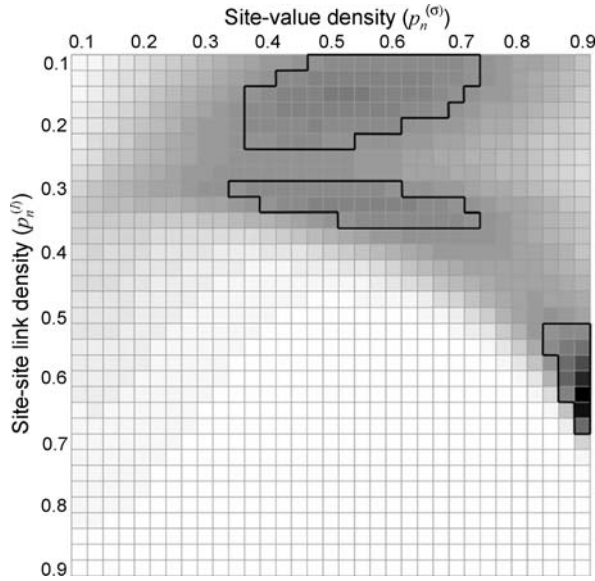
The second (link-update) function of the pair of functions in Eq. (34) is thus given by

$$G_{\text{RDA}} \rightarrow p_{n+1}^{(l)} = p_n^{(l)} \cdot (1 - P_n^a) + (1 - p_n^{(l)}) \cdot P_{D=2} \cdot P_n^a, \quad (43)$$

where, assuming that two sites,  $i$  and  $j$ , are not themselves connected,  $P_{D=2}$  = probability that there exists at least one site  $k$ , such that  $D_{ik} = D_{jk} = 1$ , which implies that  $P_{D=2} = 1 - \text{Prob}(\text{there is no such } k) = 1 - (1 - [p_n^{(l)}]^2)^{N-2}$ ; and  $P_1$  and  $P_2$  are defined in Eqs. (41) and (42).

A *structural equilibrium* is established when  $p_{n+1}^{(l)} \approx p_n^{(l)}$ , which happens when the average number of new connections is equal to the average number of link deletions:  $p_n^b \cdot \langle N_{\text{nn}} \rangle = P_n^a \cdot \langle \text{deg} \rangle$ , where  $\langle \text{deg} \rangle = p_n^{(l)}(N-1)$  is the average *degree*, and  $\langle N_{\text{nn}} \rangle$  is the average number of *next-nearest neighbors*. For SDCA rules that naturally tend to produce graphs with minimal site value and structural





Structurally Dynamic Cellular Automata, Figure 16

Density-plot of  $\gamma_{c:d}$  for an **OT decoupler** rule:  $\{(0, 0), (1, 1), (1, 2), (2, 2)\}$ ; an **OT coupler** rule:  $\{(1, 1)\}$ ;  $0.1 \leq p_n^{(\sigma)} \leq 0.9$ ; and  $0.1 \leq p_n^{(l)} \leq 0.9$ ; the rectangular area highlighted in black denotes the “equilibrium boundary” that separates regions of growth and decay

correlations, the predicted ratio of RDA link creations to deletions,  $\gamma_{c:d} \equiv P_n^b \cdot \langle N_{nn} \rangle / P_n^a \cdot \langle \text{deg} \rangle$ , may be used to predict qualitatively how the graphs will evolve. Since the average number of pairs of sites a distance  $D = 2$  apart  $= \binom{N}{2} \cdot P_{D=2} = \langle N_{nn} \rangle \cdot N/2$ , we find that:

$$\gamma_{c:d} = \frac{P_n^b}{P_n^a} \cdot \frac{(1 - p_n^{(l)})}{p_n^{(l)}} \cdot \left\{ 1 - \left( 1 - [p_n^{(l)}]^2 \right)^{N-2} \right\}. \quad (44)$$

$\gamma_{c:d}$  is also implicitly a function of site-value density, since  $p_n^{(\sigma)}$  appears in both  $P_n^a$  and  $P_n^b$ , defined in Eq. (39).

Figure 16 shows a grayscale density-plot of  $\gamma_{c:d}$  for an **OT decoupler** rule:  $\{(0, 0), (1, 1), (1, 2), (2, 2)\}$ ; an **OT coupler** rule:  $\{(1, 1)\}$ ;  $0.1 \leq p_n^{(\sigma)} \leq 0.9$ ; and  $0.1 \leq p_n^{(l)} \leq 0.9$ . Areas that are close to *white* represent combinations of  $(p_n^{(\sigma)}, p_n^{(l)})$  for which  $\gamma_{c:d} \ll 1$ , and which therefore predict “decay”; areas that are close to *black* represent combinations of  $(p_n^{(\sigma)}, p_n^{(l)})$  for which  $\gamma_{c:d} \gg 1$ , and predict “growth”; the rectangular area highlighted in black denotes the “equilibrium boundary” that separates regions of growth and decay.

### Related Graph Dynamical Systems

The original SDCA model [28] represents *one* (albeit not entirely arbitrary) approach to dynamically coupling site values  $(\{\sigma_i\})$  and topology  $(\{l_{ij}\})$ , of the normally quies-

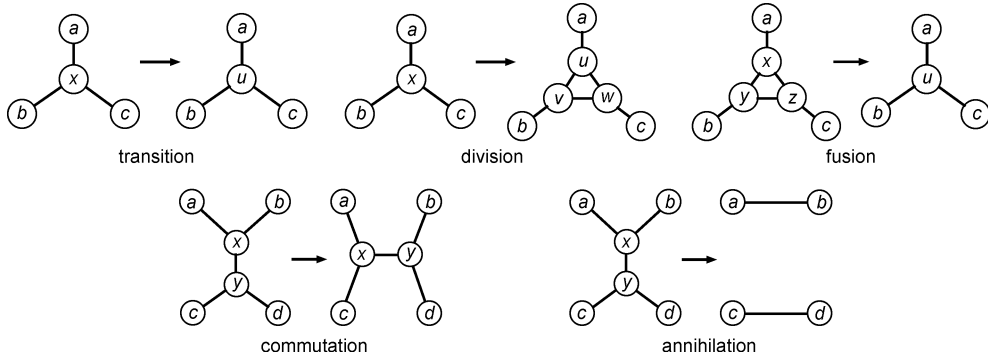
cent lattice. Since this model was primarily introduced as a general tool to explore self-organized emergent *geometries*,  $\sigma$  values are an integral dynamic component only because SDCA’s original rules were conceived to generalize conventional CA rules, not replace them. Moreover, SDCA’s link rules are, by design, close analogs of their conventional-CA brethren; this is the reason why SDCA’s  $\psi$  and  $\omega$  rules assume the familiar **T** and **OT** (and related **RT**) forms, as defined in Sect. “The Basic Model”. Indeed, while the preceding sections of this article have introduced several generalizations – such as the addition of probabilistic rules, reversibility and memory – in each case, the basic *form* of the rules (as defined in Eqs. (11), (12), and (13)) has remained essentially the same. However, just as for conventional CA, an almost endless variety of different *kinds* of rules can in principle be defined; including rules that alter the geometry but are *not* functions of the  $\sigma$  states. In this section, we look at two illustrative examples of SDCA-like dynamical systems: one that uses coupled  $\sigma$ - $\ell$  rules, and another whose rules depend only on *topology*.

### Graph Rewriting Automata

Tomita, Kurokawa, and Murata [67,68,69,70,71] have recently introduced *graph rewriting automata* (abbreviated, GRA), in which both links and (the number of) sites are allowed to change. Motivated by CA models of self-reproduction, Tomita et al suggest that fixed, two-dimensional lattices – used as static backdrops to most conventional models – are unnecessarily restrictive for describing self-reproductive processes. They cite, as an example, the inability of conventional CA to describe biological processes (such as embryonic development) that must unfold in a finite closed space; once the underlying space of the CA is defined at the start, however large (and sometimes deliberately assumed *infinite*), its size remains the same throughout the development. This not only makes it hard to model the typically growing need that developing organisms have for *space*, but makes it impractical even to provide some room for avoiding overlaps between the original and daughter patterns [67].

Motivated by these, and other issues related to computation, Tomita et al. [67,69] introduce GRA, which is a form of graph grammar [20]. At first glance, GRA appear superficially similar to SDCA, at least in the sense that they both dynamically couple site values with topology. However, the transition rules are very different, and – in GRA’s case – two properties hold that are not true for SDCA systems: (1) *all sites have exactly three neighbors at all times* (which is the minimum number of neighbors that yield nontrivial graphs [67]), and (2) *multiple links*





Structurally Dynamic Cellular Automata, Figure 17

Graphical representations of the actions of the GRA rules defined in Eq. (45). (Reproduced from [69] with permission)

are allowed to exist between any two sites. The authors claim that the 3-neighbor restriction not only does *not* constrain the space of emergent geometries (an observation that is echoed by Wolfram [75]; see Subject. “**Network Automata**” below) but has the added benefit of allowing the rules to be expressed in a regular form: each rule is defined by a rule *name* and, at most, six *symbols* for its argument:

( $\sigma$  rules) :

$$\left\{ \begin{array}{l} \text{transition}(x, a, b, c) \rightarrow (u, a, b, c), \end{array} \right.$$

(site rules) :

$$\left\{ \begin{array}{l} \text{division}(x, a, b, c) \rightarrow (u, v, w, a, b, c), \\ \text{fusion}(x, y, z, a, b, c) \rightarrow (u, a, b, c), \end{array} \right.$$

(link rules) :

$$\left\{ \begin{array}{l} \text{commutation}(x, y, a, b, c, d) \rightarrow (x, y, a, b, c, d), \\ \text{annihilation}(x, y, a, b, c, d) \rightarrow (a, b, c, d), \end{array} \right.$$

(45)

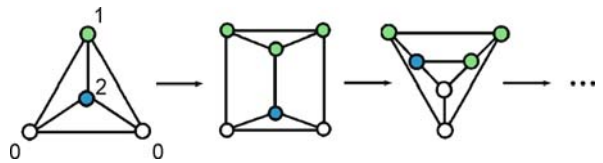
where  $x, y$  and  $z$  denote the  $\sigma$  values of the center sites *before* undergoing a structural change;  $u, v$  and  $w$  denote the  $\sigma$  values of the center sites *after* the structural change; and  $a, b, c$ , and  $d$  denote the states of the *neighboring sites*. The ordering is unimportant, so long as a given string can be obtained from another by cyclic permutation, otherwise the strings are different; i. e.,  $(a, b, c)$  is both topologically and functionally equivalent to  $(b, c, a)$ , but  $(c, b, a)$  is different. The action of  $\sigma$ , value, and links is graphically illustrated in Figure 17.

By convention, the GRA algorithm is applied in two steps: (1) site rules (*transition*, *division* and *fusion*) are executed first, and at all subsequent even time steps, followed by (2) link rules (*commutation* and *annihilation*), executed at odd steps.

In the event that multiple rules are simultaneously applicable – such as might happen, for example, if the rules include more than one division, or fusion, for the same lefthandside argument in their expressions (in Eq. (45)) – the order in which the rules are applied is determined by an a priori priority ranking. Also, since applying either *commutation* or *annihilation* rules to adjacent links yields inconsistency, whenever a local context arises in which this might happen, the application of these rules is temporarily suppressed. (This is done by sweeping through the link set *twice*: on the first pass, a temporary flag is set for each link that satisfies a rule condition; on the second pass, the link rule is applied *if and only if* the four neighboring links did not raise flags during the first pass.)

Figure 18 shows the first few steps in applying one *division* and two *commutation* rules to a simple initial graph. (Kohji Tomita provides several movies of GRA evolutions on his website: <http://staff.aist.go.jp/k.tomita/ga/>)

Tomita, Kurokawa, and Murata [67,68,69,70,71] report a variety of emergent behaviors, including (1) *arbitrary resolution* (because GRA rules effectively allow an arbitrary number of sites to “grow” out of any initial structure, these systems define their own “boundary conditions” and graphs with arbitrary resolution are possible); (2) *repetitive structures*, in which some geometrical subset of an initial graph is reproduced, indefinitely, and continu-



Structurally Dynamic Cellular Automata, Figure 18

Sample GRA evolution starting from the graph on the left. The rules are (see Eq. (45)): (1) *division*(1, 0, 0, 2)  $\rightarrow$  (1, 1, 1, 0, 0, 2), (2) *commutation*(1, 2)  $\rightarrow$  (1, 2), and (3) *commutation*(0, 0)  $\rightarrow$  (0, 0). (Reproduced from [69] with permission)

ously grafted onto the original structure; and (3) *self-replication*, in which both site-value and structure is replicated after  $N$  steps. In [70], Tomita et al. describe how genetic algorithms [45] may be used for automating the search for self-replicating patterns.

In [67], Tomita et al. also present the design of a self-reproducing *Turing Machine*. *Turing machines* are abstract symbol manipulating devices that mimic the basic operations of a computer. Formally, they consist of a “tape” (of indefinite length, to record data), a “head” (that reads/writes symbols on the tape, and that can move left or right), and “state transition rules” (that tell the head which new symbols to write given the current state of the tape). The tape is analogous to “memory” in a modern computer; the head is analogous to the microprocessor. A Turing machine is called “universal” if it can simulate any other Turing machine.

Tomita et al.’s [67] *Turing machine* is modeled as a ladder structure: the upper sites constitute the “tape” mechanism; the lower sites form the “tape head” that reads the tape; both ends of the ladder are single sites that define “end of tape”; and the two ends are joined to form a loop. Although the tape is initially finite, the ladder can grow to arbitrary length, as required, by using appropriate GRA rules. Tomita et al. [67] self-replicating Turing GRA consists of 20 states and 257 (2-symbol) rules. They also introduce a design for a universal Turing machine [69] that consists of 30 states and 955 rules for reproduction, and 23 states and 745 rules for computation. While self-reproducing universal Turing machines can be described using conventional CA, their expression using GRA rules are considerably more compact.

### Dynamic Graphs as Models of Self-Reconfigurable Robots

In the context of looking for self-reconfiguration algorithms that may be used to manufacture modular robots for industry, Saidani [63,64] has recently introduced a *dynamic graph calculus* that includes rules similar to those that define SDCA; but which depend only on the *topology* of (but not the  $\sigma$ -values living on) the lattice. Saidani and Piel [65] have also introduced an interactive programming environment for studying dynamic graph simulations called *Dynagraph*, and implemented in Smalltalk.

There are two basic approaches to designing modular robots: (1) to develop a set of elementary generic modules that can be rapidly assembled by humans to form robots that solve a specific problem, and (2) to design a set of (otherwise identical) primitive components that can adaptively reconfigure themselves. Focusing on the latter approach, Saidani [64] formally reinterprets modu-

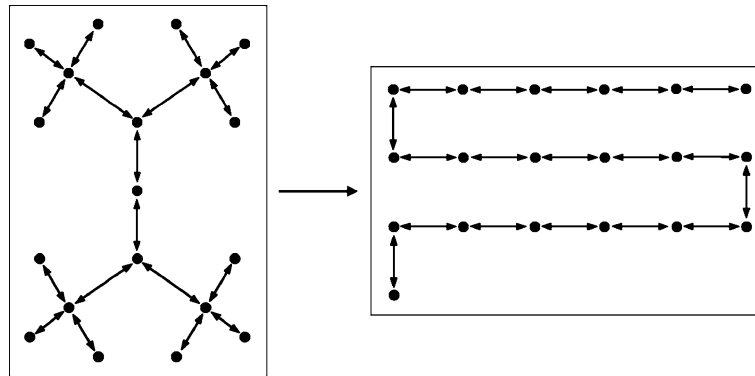
lar “robots” to mean modular *networks*; and proceeds to model adaptive robotic self-reconfigurations as a class of recursive graph dynamical systems. In contrast to other related dynamic graph models [18,25], the “modules” (or subgraphs) of Saidani’s model use local knowledge of their neighborhood topology to collectively evolve to some goal configuration. Although the dynamics transforms the global state, the evolution remains strictly decentralized, and individual modules do not know the (desired) final state.

Apart from restricting the dynamics to topology *alone* (indeed, none of the sites harbor information states of any kind), Saidani [63,64,65] further assumes that (1) connections between sites are *directional* (both *to*- and *from*-links may coexist between the same two modular components); (2) “active” sites reconfigure their local neighborhood by accepting, keeping, or removing their adjacent links according to rules that are functions of their current topology (defined as a given sites’ current local neighborhood and the current neighborhood of its neighbors: a site only knows about its own *in*- and *out*-degree, which can obviously be computed from its local topology, and the *in*- and *out*-degrees of its nearest neighbors); (3) a site controls its *outgoing* links (and can connect or disconnect any outgoing links), but cannot sever incoming connections; (4) sites must maintain at least one link throughout an evolution (so that the graph remains connected); and (5) all sites are equipped with the same set of rules.

As in conventional CA and the basic SDCA model, the “reconfiguration” proceeds synchronously throughout the graph. The decision process includes an innate stochastic element: in the event that there is a rule that specifies that a site is to establish a link to a neighbor of one of its neighbors, but all neighboring sites have the same degree (which is the only dynamical discriminant), the neighbor with which a new link will be forged is selected at random.

As a concrete example, Saidani [64] presents a *tree-to-chain* algorithm that evolves an initial “tree” graph to a linear chain of linked sites (see Fig. 19). While we do not reproduce the full algorithm here, it is essentially a case-driven list of rules of the form *if condition  $C_1$  (and condition  $C_2, \dots$  and condition  $C_n$ ) then connect (or disconnect) site  $i$  to (from) the  $n$ th neighbor of  $i$ ’s neighbor,  $v$* . For example, an explicit “rule” might be: *if  $1 \leq \deg^-(i) \leq 2$  and  $\deg^+(i) = 1$  and  $|\tau(i)| = 2$  then link  $i$  to a neighboring site  $j$  that has  $\deg^-(j) = 0$* , where  $\deg^-(i)$  and  $\deg^+(i)$  are the *in*- and *out*-degrees of site  $i$ , and  $\tau(i)$  is the total number of sites to which  $i$  is currently linked (with either incoming or outgoing links).

Conceptually, the details of Saidani’s rules are less important than what the unfolding process represents as



Structurally Dynamic Cellular Automata, Figure 19

Schematic illustration of a *tree* topology reconfiguring itself into a linear *chain* using a set of case-based “if-then” topology rules defined in [64]; see text for details

a whole. An initial graph – which we recall is to be viewed as a distillation of a “modular robot” – is transformed, by the individual sites (or parts of the robot), into another *desired* structure; i. e., the graph is entirely *self-reconfigured*. Though the broader reverse-engineering problem (which includes asking such fundamental questions as “How can a desired final state be mapped onto a specific cases-based list of graphical rules?”) remains, as yet, unanswered, and the *Dynagraph* work environment [65] is currently limited to experimenting only with graphs that have less than 30 sites, the basic model already represents a viable new approach to using dynamic graphs to describe self-reconfigurable robots; and is potentially more far-reaching as a general model of topologically-reconfigurable dynamical systems.

## SDCA as Models of Fundamental Physics

### Pregeometric Theories of Emergent Space-Time

Although SDCA are a natural formal extension of conventional CA – and serve as general-purpose modeling tools – their conception was originally motivated by fundamental physics; specifically, by a search for models of self-organized emergent discrete space-time [42]. “Space acts on matter, telling it how to move; ... matter reacts back on space, telling it how to curve”, which is the central lesson of Einstein’s *general relativity*, as explained by Misner, Thorne and Wheeler in their classic text on Gravitation [44]. Wheeler [73] has been a particularly eloquent spokesman for the need to search for what he calls a *pre-geometry*, or a set of basic elements out of which what we normally think of geometry is built, but which are themselves devoid of a specific dimensionality: “Space-time ... often considered to be the ultimate continuum of physics, evidences nowhere more clearly than at big bang and at col-

*lapse that it cannot be a continuum. Obliterated in those events is not only matter, but the space and time that envelope that matter ... we are led to ask out of what ‘pregeometry’ the geometry of space and spacetime are built*”. Wheeler has also proposed the idea that particles be viewed as geometric disturbances of spacetime, called *geometrodynamic excitons*.

A priori, SDCA appear tailor-made for describing pregeometric theories of space-time. Since in SDCA, lattice and local  $\sigma$ -values are explicitly coupled, and geometry and value configurations are treated on an approximately equal footing, SDCA is certainly at least formally consistent with Einstein’s *geometrodynamic* credo. The structure is altered locally as a function of individual site neighborhood value-states and geometries, while local site-connectivity supports the site-value evolution in exactly the same way as in conventional CA models defined on random lattices. The microphysical view of physics that emerges from this construction is one in which a fundamentally discrete pregeometry continually evolves in time as an amorphous structure but with a globally well-defined dimensionality. Particles are constructs of that amorphous structure and can be viewed as locally persistent substructures – i. e. geometrical or topological solitons – with dimensions that differ from the surrounding value. Just as “value structure” solitons are ubiquitous in conventional CA models [32,74], “link structure” solitons might emerge in SDCA; physical particles would, in such a scheme, be viewed as *geometrodynamic* disturbances propagating within a dynamic lattice.

Of course, speculation regarding the ultimate constituents of matter and space-time date back at least as far as 500 BC when the philosopher Democritus mused on the idea that matter is made of indivisible units separated by void. Since then there have been countless attempts, with

varying degrees of success, to fashion an entirely discrete theory of nature. We limit our discussion to a short survey of some recent work that centers on ideas that are either direct outgrowths of, or are otherwise conceptually related to, SDCA models. (A short history of pregeometric theories appears in chapter twelve of Ilachinski [32]).

One of the earliest proponents of pregeometry is Zuse [76], who speculated on what it would take for a CA-like universe to sustain “digital particles” on a cellular lattice. He focused on two main problems: (1) *How does the universe’s observed isotropy arise from a CA’s (Euclidean, hexagonal, etc.) anisotropy?*, and (2) *What is the information content of a physical particle?* As an answer to the first question, Zuse suggests ...

“... variable and growing automata. Irregularities of the grid structure are a function of moving patterns, which is represented by digital particles. Now, not only certain values are assigned to the single crosspoints of the grid in the concept of the cellular automaton which are interrelated and sequencing each other, but also the irregularities of the grid are itself functions of these values of the just existing interlinking network. One can imagine rather easily that in such a way the interdependence of mass, energy, and curvature of space may logically result from the behavior of the grid structure.”

Jourjine [33] generalizes Euclidean lattice field theory on a  $d$ -dimensional lattice to a cell complex. Using homology theory to replace points by cells of various dimensions and fields by functions on cells, he develops a formalism that treats space-time as a dynamical variable and describes the change in the dimension of space-time as a phase transition.

Kaplunovsky and Weinstein [34] develop a field-theoretic formalism that treats the topology and dimension of the spacetime continuum as dynamically generated variables. Dimensionality is introduced out of the characteristic behavior of the energy spectrum of a system of a large number of coupled oscillators.

Dadic and Pisk [13] introduce a self-generating discrete-space model that is based on the local quantum-mechanics of graphs. Just as in SDCA, Dadic and Pisk’s spatial structure is discrete but not static; it is fundamentally amorphous and evolves in time. Though the metric is essentially the same one used to define SDCA (i. e.,  $D_{\text{effec}}$ ), it is generalized to unlabeled graphs by referring to the topological description of the node positions rather than their arbitrary labels. Though their “graph dynamics” differs from what is used by SDCA (and uses a symmetrized Fock space that is local in terms of their graph metric, where “Fock space” is a Hilbert space used to describe quantum states with a variable, or unspecified, number of particles,

and is made from the direct sum of tensor products of single-particle; or, in this case, single-graph, Hilbert spaces) it shares two important properties with SDCA: (1) interactions depend only on the local properties of the graph, and (2) interactions induce only minimal changes to the local metric function. An important consequence of their theory is that the dimension of a graph is a scale dependent quantity that is generated by the dynamics.

**Combinatorial Space-Time** Hillman [27] introduces a *combinatorial space-time*, which he defines as a class of dynamical systems in which finite pieces of spacetime contain finite amounts of information. Spacetime is modeled as a combinatorial object, constructed by dynamically coupling copies of finitely many types of certain allowed neighborhoods. There is no *a priori* metric, and no concept of continuity, which is expected to emerge on the macroscale.

The construction (and evolution) of spaces proceeds in three steps: (1) define a set  $X$  of combinatorial  $n$ -dimensional spaces (examples are conventional CA graphs, graphs with directional links, or some other kind of embedded symmetry); (2) define a set of local, invertible primitive maps  $T: X \leftrightarrow Y$  between pairs of space sets, such that the maps do not all commute with one another (for example, a simple renaming of the sites or links gives an invertible, local map); (3) generate an arbitrary set of local invertible graph transformations by composing primitive maps with one another. Since the primitive maps are deliberately chosen so that they do not all commute, the act of composition yields infinitely many nontrivial transformations. The orbits  $\{T^z(x) \mid z \in \mathbb{Z}\}$  (for each space  $x$  in  $X$ ) are  $(n+1)$ -dimensional combinatorial spacetimes; which include reversible CA and SDCA-like networks in which geometry evolves locally over time. Formally, Hillman uses matrices of nonnegative integers, directed graphs, and symmetric tensors to describe these systems, so that local equivalences between space sets are generated by simple matrix transformations. Concrete examples of dynamic combinatorial space-time graphs are given in [27].

### Structurally Dynamic Disordered Cellular Networks

As an explicit example of how dynamic graphs can be used to model pregeometry, consider *structurally dynamic disordered cellular networks* (abbreviated, SDDCN), recently introduced by Nowotny and Requardt [53,54,55,58,59,60]. SDDCN are a class of models closely related to SDCA but developed explicitly to describe a discrete, dynamic spacetime fundamental physics. The main difference between the two models is that whereas link connections in SDCA

are strictly local, SDDCN are capable of generating both local and *translocal* links.

In contrast to more mainstream high-energy theories of fundamental physics (which are dominated by string theory and/or loop quantum gravity, both of which assume a certain level discretization at the Planck scale, but assume that a discrete space-time emerges from an underlying *continuum* physics), SDDCN takes a *bottom-up* approach. SDDCN assumes that there is underlying dynamic, discrete and highly erratic network substratum that consists of (on a given scale) irreducible mutually interacting agents exchanging information via primordial channels (links). The known continuum structures are expected to emerge on a macroscopic (or, mesoscopic) scale, via a sequence of coarse graining and/or renormalization steps.

Like SDCA, SDDCN are defined on arbitrary graphs,  $G$ , initially defined by a specified set of sites and links. Both sites and links are allowed to take on values. Site values,  $\sigma_i$  (which represent a primitive “charge”), are taken from some discrete set,  $q \cdot \mathbb{Z}$ , where  $q$  is a discrete *quantum of information*; link states assume the values  $J_{ij} \in \{-1, 0, +1\}$ , and represent an elementary coupling. The  $J_{ij}$  are equivalent to SDCA’s  $l_{ij}$ , but take on *three* values rather than two. Heuristically,  $J_{ij}$  represent directed edges pointing either from site  $i$  to  $j$  (if  $J_{ij} = 1$ ), or from  $j$  to  $i$  (if  $J_{ij} = -1$ ); or, in the case of  $J_{ij} = 0$ , the absence of a link. At each time step (representing an elementary quantum of time), an elementary quantum  $q$  is transported along each existing directed link in the indicated direction. As for SDCA, SDDCN dynamically couples site values to links.

Nowotny and Requardt [53] introduce two network models: one in which connected sites that have very different internal states typically lead to large local fluctuations (= SDDCN<sub>1</sub>), and another in which sites with similar internal states are connected (= SDDCN<sub>2</sub>):

SDDCN<sub>1</sub>

$$\Leftrightarrow \begin{cases} \sigma_i^{t+1} = \sigma_i^t + \sum_j J_{ij}^t, \\ J_{ij}^{t+1} = \text{sign}(\Delta\sigma_{ij}) \\ J_{ij}^{t+1} = 0, \end{cases} \quad \text{for } \begin{cases} |\Delta\sigma_{ij}| \geq \lambda_2, \text{ or} \\ |\Delta\sigma_{ij}| \geq \lambda_1 \wedge J_{ij}^t \neq 0, \\ \text{otherwise,} \end{cases}$$

SDDCN<sub>2</sub>

$$\Leftrightarrow \begin{cases} \sigma_i^{t+1} = \sigma_i^t + \sum_j J_{ji}^t, \\ J_{ij}^{t+1} = \text{sign}(\Delta\sigma_{ij}) \\ J_{ij}^{t+1} = J_{ij}^t, \\ J_{ij}^{t+1} = 0, \end{cases} \quad \text{for } \begin{cases} 0 < |\Delta\sigma_{ij}| < \lambda_1, \text{ or} \\ 0 < |\Delta\sigma_{ij}| < \lambda_2 \wedge J_{ij}^t \neq 0, \\ \text{for } \Delta\sigma_{ij} = 0, \\ \text{otherwise,} \end{cases} \quad (46)$$

where  $\Delta\sigma_{ij} = \sigma_i^t - \sigma_j^t$ , and  $\lambda_2 \geq \lambda_1 \geq 0$ . Since SDDCN is intended to model pregeometric dynamics, Nowotny and Requardt [53] caution that the  $t$  parameter that appears in these equations must not to be confused with the “true time” that (they expect) emerges on coarser scales. In keeping with its physics-based motivation, SDDCN’s dynamical laws depend only on the relative differences in site values, not on their absolute values. Indeed, charge is nowhere either created or destroyed, so that SDDCN conserves global “charge”:  $\sum_i \sigma_i^t = \text{constant}$ , where the arbitrary constant can be set to zero.

Both models start out initially on a simplex graph with  $N \sim 200$  nodes, so that the maximum number of possible links is  $N(N-1)/2$ . The initial  $\sigma$ -seed consists of a uniform random distribution of values scattered over the interval  $\{-k, -k+1, \dots, k-1, k\}$ , where  $k \sim 100$ . The initial values for link states,  $J_{ij}^{t=0}$ , are selected from  $\{-1, 1\}$  with equal probability; i. e., the initial state is a maximally entangled nucleus of nodes and links. Nowotny and Requardt [55] state that “... in a sense, this is a scenario which tries to imitate the big bang scenario. The hope is, that from this nucleus some large-scale patterns may ultimately emerge for large clock-time”. For most properties (other than the  $\langle \sigma \rangle_t$  and  $\sum_i |\sigma_i^{t+1} - \sigma_i^t|$ , which are both equal to zero by construction), the average over the width of the initial vertex state distribution, taken over  $\lambda_1$  and  $\lambda_2$ , specific realizations of initial conditions, and time, depend linearly on network size.

We summarize Nowotny’s and Requardt’s [53,55] findings, culled from extensive numerical experiments: (1) the appearance of very short limit cycles in SDDCN<sub>1</sub> (period 6 and multiples of 6, with the longest having period 36 on a network of size  $N = 800$ ), (2) Much longer limit cycles and transients in SDDCN<sub>2</sub>, both of which appear to grow approximately exponentially, (3) structurally, SDDCN<sub>1</sub> evolve from almost fully connected simplex networks to more sparse connectivities with increasing  $\lambda_{1/2}$ ; there is a regime in which few vertices with very high degree coexist with many vertices with a low degree; for large (around  $\lambda_1 \approx 60$ ; for large  $\lambda_{1/2}$ , the graph eventually breaks apart and all nodes become isolated; (4) for SDDCN<sub>2</sub>, nodes typically have zero degree small  $\lambda_{1/2}$ , and links become increasingly dense as  $\lambda_{1/2}$  increase; the degree distribution is generally broad and remains so for large  $\lambda_{1/2}$  (the authors also note observing multiple local maxima of the distributions in a wide range of  $\lambda_{1/2}$  values); (5) for SDDCN<sub>1</sub>, there is an abrupt phase-transition in the temporal fluctuations of vertex degrees (defined as  $\text{deg}_i(t+1) - \text{deg}_i(t)$ ) from a state in which there are essentially no fluctuations (“frozen network”) to one with strong fluctuations (“liquid network”); (6) the distribu-



tion of site values is strongly bimodal for  $62 \leq \lambda_1 \leq 85$  for  $\text{SDDCN}_1$  (while  $\text{SDDCN}_1$  distributions are not bimodal, the width of the site value distributions for different values of  $\lambda_1$  appears modulated).

From a fundamental physics perspective, the most interesting class of behaviors of  $\text{SDDCN}$  involves emergent dimensionality. Nowotny and Requardt [55] argue that since the continuum is a self-organized dynamic structure that emerges in the limit of large  $N$  and  $t$ , the most useful measure of “dimension” cannot be purely local (as in the case of *effective dimensionality*,  $D_{\text{effec}}$ , used for describing SDCA systems). Rather, it must be an intrinsically *global property*; one that is independent of any arbitrary embedding dimension, and one that can take on relatively stable values in the *whole* (to characterize effective system-wide characteristics), while simultaneously being relatively impervious to otherwise rapidly changing structural changing taking place on the microscale. Toward this end, Nowotny and Requardt [53] define the upper (and lower) scaling dimensions,  $D_S^U(i)$  (and  $D_S^L(i)$ ), with respect to site  $i$ :

$$\begin{aligned} D_S^U(i) &= \limsup_{r \rightarrow \infty} \frac{\ln \beta(i, r)}{\ln r}, \\ D_S^L(i) &= \liminf_{r \rightarrow \infty} \frac{\ln \beta(i, r)}{\ln r}, \end{aligned} \quad (47)$$

and the upper (and lower) connectivity dimensions,  $D_C^U(i)$  (and  $D_C^L(i)$ ), with respect to site  $i$ :

$$\begin{aligned} D_C^U(i) &= \limsup_{r \rightarrow \infty} \frac{\ln \partial \beta(i, r)}{\ln r}, \\ D_C^L(i) &= \liminf_{r \rightarrow \infty} \frac{\ln \partial \beta(i, r)}{\ln r}, \end{aligned} \quad (48)$$

where  $\beta(i, r) = \# \text{ sites } j \mid D_{ij} \leq r$ , and  $\partial \beta(i, r)$  is the number of sites on the *surface* of the  $r$ -sphere. When the upper and lower limits coincide, we have the *scaling dimension* ( $= D_S$ ) and the *connectivity dimension* ( $= D_C$ ), respectively.  $D_S$  is related to well known dimensional concepts in fractal geometry;  $D_C$  is a more physical measure that describes how the graph is connected, and thus how sites may potentially influence one another [55]. Preliminary research [53] suggests that under certain conditions, behavior resembling a structural phase transition to states with stable internal (and/or connectivity) dimensions is possible.

### Network Automata

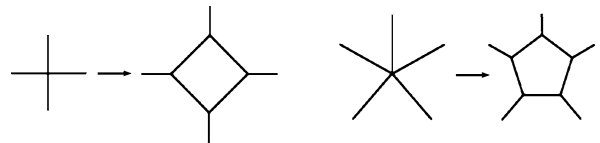
Stephen Wolfram devotes chapter nine of his Opus – *A new kind of science* (abbreviated, NKS) [75] – to applying CA to fundamental physics; and speculates on ways in

which space may be described using a dynamic network. The central, overarching theme of NKS is that “simple” programs often suffice to capture complex behaviors.

The bold claim made in chapter nine of NKS is that, on an even more fundamental level, what underlies *all the laws of physics*, as we currently understand them, is a simple CA-like program, from which, ultimately, all the phenomenologically observed complexity in the universe naturally emerges. As for the specific forms such a “program” may take, Wolfram’s intellectual point of departure echoes that of other proponents of a discrete dynamic pregeometric theory:

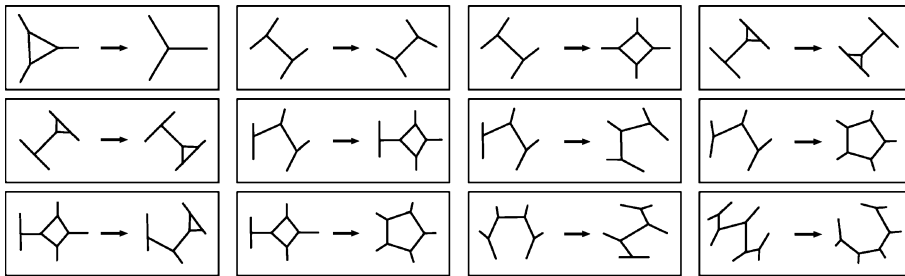
“... cellular automata ... cells are always arranged in a rigid array in space. I strongly suspect that in the underlying rule for our universe there will be no such built-in structure. Rather ... my guess is that at the lowest level there will just be certain patterns of connectivity that tend to exist, and that space as we know it will then emerge from these patterns as a kind of large-scale limit”.

Wolfram introduces his *network automata* (abbreviated, NA) with these basic assumptions (see additional notes in NKS [75] on the evolution of networks: pp. 1037–1040): (1) features of our universe emerge solely from properties of space, (2) the underlying model (and/or “rules”) must contain only a minimal underlying geometric structure, (3) the individual sites of emergent graphs must not be assigned any intrinsic position, (4) sites are limited to possessing purely *topological* information (that defines the set of sites to which a given site is connected), (5) incoming and outgoing connections need not be distinguished, and (6) all sites have exactly the same total number of links to other sites (which is assumed equal to *three*). This last assumption – which is essentially the same one made by Nowotny and Requardt [53] as the basis of their  $\text{SDDCN}$  model; see Subsect. “Structurally Dynamic Disordered Cellular Networks” above – does not lead to any loss of generality. With two connections, only very trivial graphs are possible; and it is easy to show that any site with more than three links can always be redefined, locally, as a collection of sites with exactly three links each (see Fig. 20).



Structurally Dynamic Cellular Automata, Figure 20

Illustration of how sites that have more than three links can always be redefined as a set of sites with exactly three links each



Structurally Dynamic Cellular Automata, Figure 21

Examples of planarity-preserving network substitution rules. (Reproduced from [75] with permission)

Wolfram [75] gives several concrete examples of evolving graphs (as models of pregeometry), the dynamics of which are prescribed by a set of *substitution* rules; i. e., explicit lists of the topological configurations (of sites and links) that are used to replace (at time  $t + 1$ ) specific local configurations (as they appear at time  $t$ ). However, in contrast to SDCA rules, Wolfram’s substitution rules are strictly *topological*; no site-value information is used. Also, the number of sites in the graph can change as the graph evolves; where, in SDCA, the number remains constant.

Figure 21 shows examples of rules in which specific clusters of sites are replaced with other clusters of sites. While the rules shown in the figure share the property that they all preserve planarity, there is no particular reason for imposing such a restriction; in fact, rules that generate non-planarity are just as easy to define. Wolfram speculates (pp. 526–530 in [75]) that “particle states” may be defined as mobile *non*-planar subgraphs that persist on an otherwise planar, but *randomly* fluctuating topology. Reversible versions of these rules may also be constructed, by associating a “backward” version with each “forward” transformation.

Some care must be taken while both *defining* and *applying* these rules consistently. For example, if a cluster of sites contains a certain number of links at  $t$ , one is not permitted to define a rule that replaces that cluster with another one that has a different number of connections. Another restriction is that rules must be independent of orientation; that is, if a candidate rule requires identifying the specific links (of, say, an otherwise topologically symmetric  $n$ -link local subgraph) before activating a desired substitution, that rule is likewise forbidden. However, even with these restrictions, a large number of rules are still possible. For example, 419 distinct rules may be defined for clusters with no more than five sites.

In applying network rules, one cannot simply simultaneously replace all pertinent subgraphs with their replacements, since, in general, two or more subgraphs with the same topology may overlap somewhere within the net-

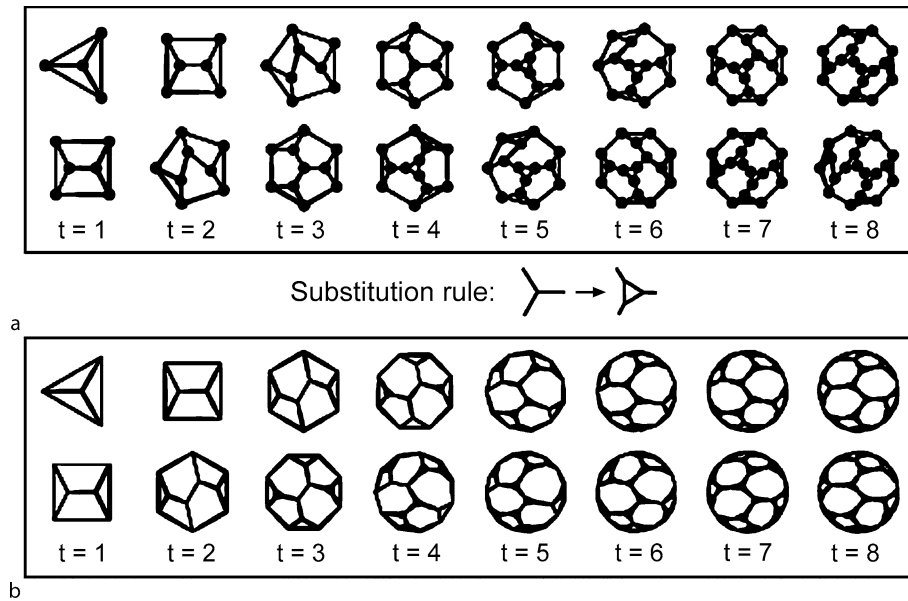
work. Since there is no *priori*, or universally consistent, way of ordering the subgraphs, *meta*-rules must be imposed to eliminate any possible ambiguities. For example, one method (*m1*) is to restrict replacements to a single subgraph per time step, selecting the subgraph whose replacement entails the minimal change to all recently updated sites. Another method (*m2*) is to allow all possible nonoverlapping replacements, while ignoring those that overlap. Wolfram reports that, although the second method obviously produces larger graphs in fewer steps, the two methods generally produce qualitatively similar structures.

Figure 22 traces the first few steps in the evolution of a simple graph under the action of a single substitution rule (defined at the center of the figure). Figures 22a and 22b show the results of applying this rule using methods *m1* and *m2*, respectively. In each case, the top row shows the form of the network before the substitution takes place at that step, and the bottom row shows the network that results from the substitution. The subgraph (or subgraphs, in Fig. 22a) involved in the replacement is highlighted at both top and bottom.

Wolfram also suggests that analogs of *mobile automata* [43] can be defined for evolving networks. By tagging a site  $i$ , say, with a “charge”,  $\sigma_i \equiv 1$ , substitution rules may be defined to replace clusters of sites around the charged site. The effect is that the charge itself appears to move, as its effective (relative) position within the network changes as the geometric dynamics unfolds. (However, Wolfram also notes – on page 1040 in [75] – that “*despite looking at several hundred cases I have not been able to find network mobile automata with especially complicated behavior*”).

### Future Directions and Speculations

Although SDCA were first introduced over two decades ago [28], much of their behavior remains unexplored. Of course, this is due largely to the difficulty of studying dy-



Structurally Dynamic Cellular Automata, Figure 22

Examples of network evolutions using the substitution rule shown at center. See text for explanation. (Reproduced from [75] with permission)

namical systems that harbor an a priori vastly larger coupled value-geometry space than the “merely” spatially-confined behavioral space of conventional CA. Only relatively recently have desktop computers become sufficiently powerful, and visualization programs adept enough at rendering multidimensional graphs [12], to make a serious study of SDCA behaviors possible. For example, the general-purpose math programs *Mathematica* (<http://www.wri.com>) and *Maple* (<http://www.maplesoft.com>) both provide powerful built-in graph-rendering algorithms to help visualize complex graphs. Standalone public-domain packages are also available; for example, AGNA [2], Net-Draw [11], and Pajek [52]. In this final section, we list several open questions and briefly speculate on possible future directions.

Because of the relative paucity of studies dedicated purely to exploring the space of emergent structures (such as Wolfram’s [74] pioneering studies of conventional CA), many (even very fundamental) questions remain open: *What kinds of geometries can arise?*, *Which subspace of the space of all possible graphs corresponds to those that are actually attainable using SDCA (and SDCA-like) rules?*, *What are the conditions for which certain geometries do, and do not, form?*, *What combinations of  $\sigma$ - and  $\ell$ -rules give rise to specific kinds of graphs?*

Other open problems include: (1) determining whether the (provisionally defined) set of class-4 rules, for which effective dimension appears to remain constant, is

*genuine*, rather than being either a long-term transient or an unintentional artifact of imposed run-time constraints; and, if this class is “real”, we obviously need to ask, *How large is it?*, and *Under what conditions does it arise?*; (2) developing SDCA as formal mathematical models, perhaps as members of a broader class of graph grammars [20,35]; and (3) finding purely geometric analogs of the solitons known to exist in conventional CA models [32,74].

This article has introduced several generalizations of the basic SDCA model, including *memory* effects (Subsect. “SDCA With Memory”), *reversibility* (Subsect. “Reversible SDCA”), *probabilistic transitions* (Subsect. “Probabilistic SDCA”), and a class of SDCA-like dynamical systems that evolve according to rules that depend only on *topology* (Subsects. “Dynamic Graphs as Models of Self-Reconfigurable Robots” and “Network Automata”). However, other possibilities abound: (1)  $\sigma$  site-variables may take on a larger range of values,  $\sigma \in \{0, 1, \dots, k-1\}$ ; (2) link variables,  $\ell_{ij}$ , may similarly take on a larger range of values,  $\ell_{ij} \in \{0, \pm 1, \pm 2, \dots, \pm m\}$  (where, say,  $\pm$  determines “directionality”, and absolute value,  $|\ell_{ij}|$ , represents either channel capacity for information flow or some other innate property); and (3) both sites and links may take on richer, and more explicitly “active”, roles of *agent-actors* [17].

Apart from these formal extensions, some obvious future applications include modeling communication and social network dynamics, studying the dynamics of plas-

ticity in artificial neural networks, designing adaptive self-reconfiguring parallel-computer networks (as well as “amorphous” computer chips), studying behaviors of gene-regulatory networks, and providing the conceptual core for fundamental pregeometric physical theories of discrete, emergent space-times.

## Bibliography

### Primary Literature

- Adamatzky A (1995) *Identification of Cellular Automata*. Taylor and Francis, London
- Applied Graph & Network Analysis software. <http://benta.addr.com/agna/>. Accessed 14 Oct 2008
- Albert J, Culik II K (1987) A simple universal cellular automaton and its one-way and totalistic version. *Complex Syst* 1:1–16
- Ali SM, Zimmer RM (1995) Games of Proto-Life in Masked Cellular Automata. *Complexity International*, vol 2. <http://www.complexity.org.au>
- Ali SM, Zimmer RM (2000) A Formal Framework for Emergent Panpsychism. In: Tucson 2000: Consciousness Research Abstracts. <http://www.consciousness.arizona.edu/tucson2000/>. Accessed 14 Oct 2008
- Alonso-Sanz R, Martin M (2006) A structurally dynamic cellular automaton with memory in the hexagonal tessellation. In: El Yacoubi S, Chopard B, Bandini S (eds) *Lecture Notes in Computer Science*, vol 4173. Springer, New York, pp 30–40
- Alonso-Sanz R (2006) The Beehive Cellular Automaton with Memory. *J Cell Autom* 1(3):195–211
- Alonso-Sanz R (2007) A structurally dynamic cellular automaton with memory. *Chaos, Solitons, Fractals* 32(4):1285–1304
- Barabasi AL, Albert R (2002) Statistical mechanics of complex networks. *Rev Mod Phys* 74:47–97
- Bollobas B (2002) *Modern Graph Theory*. Springer, New York
- Borgatti SP (2002) NetDraw 1.0: Network visualization software. Analytic Technologies, Harvard
- Chen C (2004) *Graph Drawing Algorithms*. In: *Information Visualization*. Springer, New York
- Dadic I, Pisk K (1979) Dynamics of discrete-space structure. *Int J Theor Phys* 18:345–358
- Doi H (1984) Graph theoretical analysis of cleavage pattern: graph developmental system and its application to cleavage pattern in ascidian egg. *Dev Growth Differ* 26(1):49–60
- Durrett R (2006) *Random Graph Dynamics*. Cambridge University Press, New York
- Erdos P, Renyi A (1960) On the evolution of random graphs. *Publ Math Inst Hung Acad Scie* 5:11–61
- Ferber J (1999) *Multi-Agent Systems: An Introduction to Distributed Artificial Intelligence*. Addison-Wesley, New York
- Ferreira A (2002) On models and algorithms for dynamic communication networks: the case for evolving graphs. In: 4th Recontres Francophones sur les Aspects Algorithmiques des Télécommunications (ALGOTEL 2002), Meze, France
- Gerstner W, Kistler WM (2002) *Spiking Neuron Models*. Single Neurons, Populations, Plasticity. Cambridge University Press, New York
- Grzegorz R (1997) *Handbook of Graph Grammars and Computing by Graph Transformation*. World Scientific, Singapore
- Halpern P (1989) Sticks and stones: a guide to structurally dynamic cellular automata. *Amer J Phys* 57(5):405–408
- Halpern P, Caltagirone G (1990) Behavior of topological cellular automata. *Complex Syst* 4:623–651
- Halpern P (1996) Genetic algorithms on structurally dynamic lattices. In: Toffo T, Biafore M, Leao J (eds) *PhysComp96*. New England Complex Systems Institute, Cambridge, pp 135–136
- Halpern P (2003) Evolutionary Algorithms on a Self-Organized, Dynamic Lattice. In: Bar-Yam Y, Minai A (eds) *Unifying Themes in Complex Systems*, vol 2. *Proceedings of the Second International Conference on Complex Systems*. Westview Press, Cambridge
- Harary F, Gupta G (1997) Dynamic graph models. *Math Comp Model* 25(7):79–87
- Hasslacher B, Meyer D (1998) Modeling dynamical geometry with lattice gas automata. *Int J Mod Phys C* 9:1597
- Hillman D (1995) *Combinatorial Spacetimes*. Ph D Dissertation, University of Pittsburg
- Ilachinski A (1986) *Topological life-games I*. Preprint. State University of New York at Stony Brook
- Ilachinski A, Halpern P (1987) Structurally dynamic cellular automata. Preprint. State University of New York at Stony Brook
- Ilachinski A, Halpern P (1987) Structurally dynamic cellular automata. *Complex Syst* 1(3):503–527
- Ilachinski A (1988) Computer Explorations of Self Organization in Discrete Complex Systems. *Diss Abstr Int B* 49(12):5349
- Ilachinski A (2001) *Cellular Automata: A Discrete Universe*. World Scientific, Singapore
- Jourjine AN (1985) Dimensional phase transitions: coupling of matter to the cell complex. *Phys Rev D* 31:1443
- Kaplunovsky V, Weinstein M (1985) Space-time: arena or illusion? *Phys Rev D* 31:1879–1898
- Kniemeyer O, Buck-Sorlin GH, Kurth W (2004) A Graph Grammar Approach to Artificial Life. *Artif Life* 10(4):413–431
- Krivovichev SV (2004) Crystal structures and cellular automata. *Acta Crystallogr A* 60(3):257–262
- Lehmann KA, Kaufmann M (2005) Evolutionary algorithms for the self-organized evolution of networks. In: *Proceedings of the 2005 Conference on Genetic and Evolutionary Computation*, Washington DC. ACM Press, New York
- Love P, Bruce M, Meyer D (2004) Lattice gas simulations of dynamical geometry in one dimension. *Phil Trans Royal Soc A: Math Phys Eng Sci* 362(1821):1667–1675
- Majercik S (1994) *Structurally dynamic cellular automata*. Master's Thesis, Department of Computer Science, University of Southern Maine
- Makowiec D (2004) Cellular Automata with Majority Rule on Evolving Network. In: *Lecture Notes in Computer Science*, vol 3305. Springer, Berlin, pp 141–150
- Mendes RV (2004) Tools for network dynamics. *Int J Bifurc Chaos* 15(4):1185–1213
- Meschini D, Lehto M, Piilonen J (2005) Geometry, pregeometry and beyond. *Stud Hist Phil Mod Phys* 36:435–464
- Miramontes O, Solé R, Goodwin B (1993) Collective behavior of random-activated mobile cellular automata. *Physica D* 63:145–160
- Misner CW, Thorne KS, Wheeler JA (1973) *Gravitation*. W.H. Freeman, New York
- Mitchell M (1998) *An Introduction to Genetic Algorithms*. MIT Press, Boston

46. Moore EF (1962) *Sequential Machines: Selected Papers*. Addison-Wesley, New York
47. Muhlenbein H (1991) Parallel genetic algorithm, population dynamics and combinatorial optimization. In: Schaffer H (ed) *Third International Conference on Genetic Algorithms*. Morgan Kaufman, San Francisco
48. Murata S, Tomita K, Kurokawa H (2002) System generation by graph automata. In: Ueda K (ed) *Proceedings of the 4th International Workshop on Emergent Synthesis (IWES '02)*, Kobe University, Japan, pp 47–52
49. Mustafa S (1999) *The Concept of Poiesis and Its Application in a Heideggerian Critique of Computationally Emergent Artificiality*. Ph D Thesis, Brunel University, London
50. Newman M, Barabasi A, Watts DJ (2006) *The Structure and Dynamics of Networks*. Princeton University Press, New Jersey
51. Nochella J (2006) Cellular automata on networks. Talk given at the Wolfram Science Conference (NKS2006), Washington DC, USA, 16–18 June
52. Nooy W, Mrvar A, Batagelj V (2005) *Exploratory Social Network Analysis with Pajek*. Cambridge University Press, New York
53. Nowotny T, Requardt M (1998) Dimension Theory of Graphs and Networks. *J Phys A* 31:2447–2463
54. Nowotny T, Requardt M (1999) Pregeometric Concepts on Graphs and Cellular Networks as Possible Models of Space-Time at the Planck-Scale. *Chaos, Solitons, Fractals* 10:469–486
55. Nowotny T, Requardt M (2006) Emergent Properties in Structurally Dynamic Disordered Cellular Networks. [arXiv: cond-mat/0611427](https://arxiv.org/abs/cond-mat/0611427). Accessed 14 Oct 2008
56. O'Sullivan D (2001) Graph-cellular automata: a generalized discrete urban and regional model. *Environ Plan B: Plan Des* 28(5):687–705
57. Prusinkiewicz P, Lindenmayer (1990) *The Algorithmic Beauty of Plants*. Springer, New York
58. Requardt M (1998) Cellular Networks as Models for Planck-Scale Physics. *J Phys A* 31:7997–8021
59. Requardt M (2003) A geometric renormalisation group in discrete quantum space-time. *J Math Phys* 44:5588–5615
60. Requardt M (2003) Scale free small world networks and the structure of quantum space-time. [arXiv.org:gr-qc/0308089](https://arxiv.org/abs/gr-qc/0308089)
61. Rose H (1993) *Topologische Zellaere Automaten*. Master's Thesis, Humboldt University of Berlin
62. Rose H, Hempel H, Schimansky-Geier L (1994) Stochastic dynamics of catalytic CO oxidation on Pt(100). *Physica A* 206:421–440
63. Saidani S (2003) *Topodynamique de Graphe. Les Journées Graphes, Réseaux et Modélisation*. ESPCI, Paris
64. Saidani S (2004) Self-reconfigurable robots topodynamic. In: *IEEE International Conference on Robotics and Automation*, vol 3. IEEE Press, New York, pp 2883–2887
65. Saidani S, Piel M (2004) DynaGraph: a Smalltalk Environment for Self-Reconfigurable Robots Simulation. *European Smalltalk User Group Conference*. <http://www.esug.org/>
66. Schliecker G (1998) Binary random cellular structures. *Phys Rev E* 57:R1219–R1222
67. Tomita K, Kurokawa H, Murata S (2002) Graph Automata: Natural Expression of Self-Reproduction. *Physica D* 171(4):197–210
68. Tomita K, Kurokawa H, Murata S (2005) Self-description for construction and execution in graph rewriting automata. In: *Lecture Notes in Computer Science*, vol 3630. Springer, Berlin, pp 705–715
69. Tomita K, Kurokawa H, Murata S (2006) Two-state graph-rewriting automata. *NKS 2006 Conference*, Washington, DC
70. Tomita K, Kurokawa H, Murata S (2006) Automatic generation of self-replicating patterns in graph automata. *Int J Bifurc Chaos* 16(4):1011–1018
71. Tomita K, Kurokawa H, Murata S (2006) Self-Description for Construction and Computation on Graph-Rewriting Automata. *Artif Life* 13(4):383–396
72. Weinert K, Mehnen J, Rudolph G (2002) Dynamic Neighborhood Structures in Parallel Evolution Strategies. *Complex Syst* 13(3):227–244
73. Wheeler JA (1982) The computer and the universe. *Int J Theor Phys* 21:557
74. Wolfram S (1984) Universality and complexity in cellular automata. *Physica D* 10:1–35
75. Wolfram S (2002) *A New Kind of Science*. Wolfram Media, Champaign, pp 508–545
76. Zuse K (1982) The Computing Universe. *Int J Theor Phys* 21:589–600

## Books and Reviews

- Battista G, Eades P, Tamassia R, Tollis IG (1999) *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall, New Jersey
- Bornholdt S, Schuster HG (eds) (2003) *Handbook of Graphs and Networks*. Wiley-VCH
- Breiger R, Carley K, Pattison P (2003) *Dynamical Social Network Modeling and Analysis*. The National Academy Press, Washington DC
- Dogogovtsev SN, Mendes JF (2003) *Evolution of Networks*. Oxford University Press, New York
- Durrett R (2006) *Random Graph Dynamics*. Cambridge University Press, New York
- Gross JL, Yellen J (eds) (2004) *Handbook of Graph Theory*. CRC Press, Boca Raton

## Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy

BENJAMIN A. BROOKS<sup>1</sup>, JAMES H. FOSTER<sup>1</sup>,

JEFFREY J. MCGUIRE<sup>2</sup>, MARK BEHN<sup>2</sup>

<sup>1</sup> School of Ocean and Earth Science and Technology, University of Hawaii, Honolulu, USA

<sup>2</sup> Department of Geology and Geophysics, Woods Hole Oceanographic Institution, Woods Hole, USA

## Article Outline

Glossary

Definition of the Subject

Introduction

Monitoring Motion:

Subaerial and Submarine Geodetic Methods



Data Analysis and Inversion  
 Discussion: Slow Earthquake  
 and Submarine Landslide Process  
 Future Directions: Slow Earthquakes  
 and Submarine Landslide Monitoring  
 Acknowledgments  
 Bibliography

## Glossary

**Submarine landslide** A gravitational mass failure feature on the seafloor.

**Slow earthquake** A discrete slip event that produces millimeter to meter-scale displacements identical to those produced during earthquakes but without the associated seismic shaking.

**GPS** The Global Positioning System consists of a constellation of at least 24 medium earth orbiting satellites transmitting two or more microwave frequencies for use in precise positioning.

**Seafloor geodesy** The application of geodetic methods (studies of the change in the shape of the earth's surface) applied to a submarine environment.

## Definition of the Subject

The term 'submarine landslide' encompasses a multitude of gravitational mass failure features at areal scales from square meters to thousands of square kilometers. Here, we concentrate on the large end of that spectrum, namely the submarine landslides that, when they move either in contained slip events or catastrophically, can generate surface displacements equivalent to  $> M6$  earthquakes and/or hazardous tsunamis.

The term 'slow earthquake' describes a discrete slip event that produces millimeter to meter-scale displacements identical to those produced during earthquakes but without the associated seismic shaking. Slow earthquakes, primarily associated with tectonic fault zones, have been recognized and studied with increasing frequency in the past decade largely due to the decreasing cost and proliferation of Global Positioning System (GPS) geodetic networks capable of detecting the ground motion [1,2,3]. Recently, one such GPS network on the south flank of Kilauea volcano, has recorded multiple slow earthquakes on the subaerial portion of a large landslide system that occurs primarily in the submarine environment [4,5,6]. Because the bathymetric charts surrounding the Hawaiian islands are littered with the remnants of massive, catastrophically emplaced submarine landslides (Fig. 1) it is natural to wonder if a slow-slipping submarine landslide

is a precursory stage of one that will ultimately fail catastrophically.

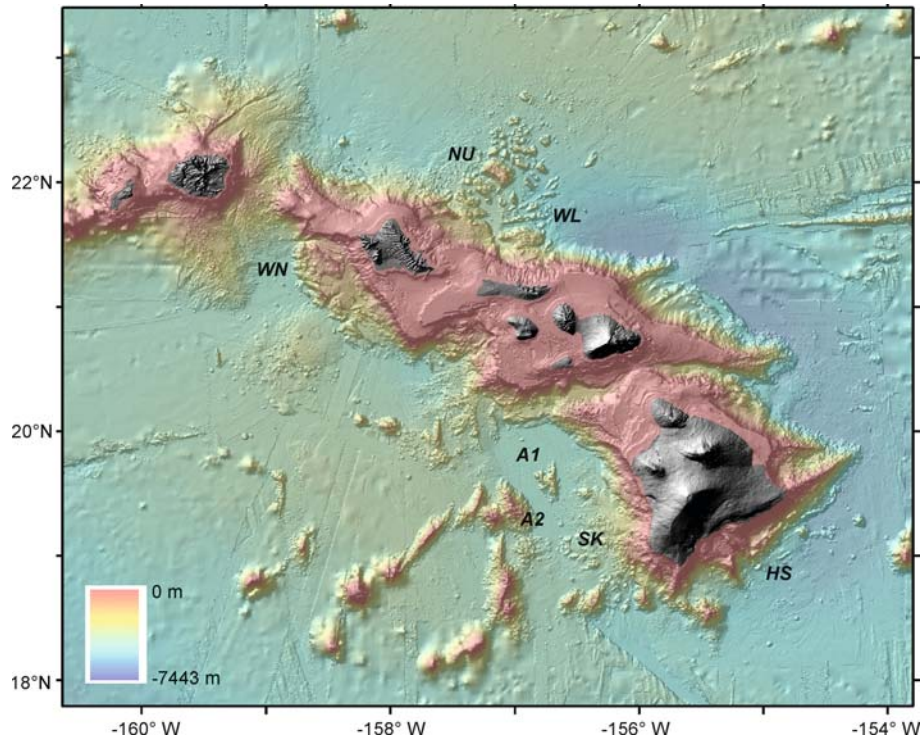
We see two principal reasons why monitoring submarine landslides and slow earthquakes associated with them is important. First, because catastrophic failure of submarine landslides can cause tsunamis they represent significant hazards to coastal zones. Understanding and monitoring how slow slip may lead to accelerated slip and catastrophic failure is, therefore, very important in terms of hazard mitigation. Second, submarine landslide systems can be some of the most active as well as spatially confined deforming areas on earth and so they represent excellent targets of study for furthering our understanding of the general fault failure process. For instance a pertinent question for which we do not yet have an answer is: are fault frictional properties homogeneous enough that the occurrence of slow earthquakes on a detachment fault plane underlying a landslide could relieve stress on the fault or do the slow earthquakes in one region load a neighboring seismogenic patch bringing it closer to a large sudden failure (i. e. an earthquake)?

While installation and operation of GPS networks on land is now relatively routine and somewhat inexpensive, the in situ monitoring of submarine landslide motion represents a significant technical challenge with accordingly higher costs. Submarine geodesy e.g. [7], however, is a nascent and rapidly evolving field with relative and absolute positioning techniques being intensely studied and developed. The near future is sure to see many advances in our monitoring and understanding of submarine landslides and slow earthquakes due to the application of submarine geodetic techniques.

## Introduction

### Submarine Landslides

The last 30 years have seen a dramatic increase in the recognition of submarine landslides world-wide, due largely to the increased prevalence and capability of swath and side-looking sonar mapping systems and systematic submarine mapping programs. For instance, the side-scan sonar mapping of the Hawaiian exclusive economic zone in the 1980s resulted in the discovery that massive submarine landslides are spatially distributed along the entire Hawaiian Ridge [8] (Fig. 1). These features are some of the largest landslides on the planet, with more than 70 attaining lengths greater than 20 km and some having lengths greater than 200 km and total volumes exceeding 5000 km<sup>3</sup>. Since then, submarine landslides associated with other volcanic islands e.g. [9,10,11], mid-ocean ridges [12], and continental margins [13] have also



**Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 1**

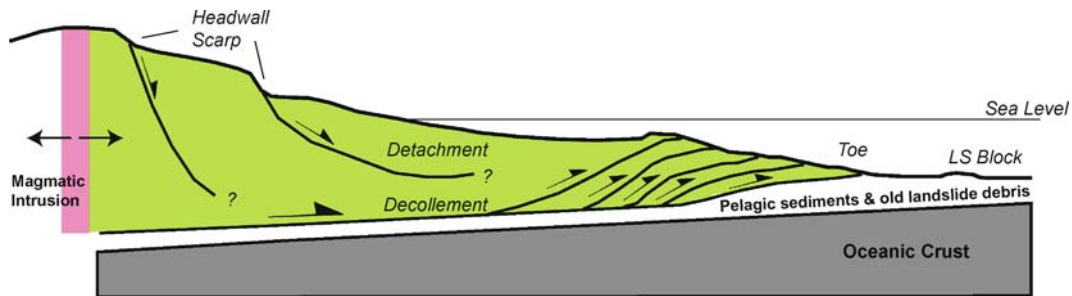
Topographic and bathymetric map of the Hawaiian Islands (data from <http://geopubs.wr.usgs.gov/i-map/i2809/>). Studied submarine landslides are indicated: HS, Hilina Slump; SK, South Kona; A1, Alika 1; A2, Alika 2; WL, Wailau; NU, Nu'uuanu; WN, Wai'anae

been studied. Of these, the slopes flanking volcanic islands, especially when they are in their steeper-sloped shield-building stage [14,15], tend to be particularly susceptible to landslide instability and so they have been the focus of much recent research in the Canary Islands [16,17] and especially in the Hawaiian Islands [18,19,20,21,22,23]. For instance, most of the Hawaiian submarine landslides are thought to be inactive, except for those on the flanks of the Big Island e.g. [21].

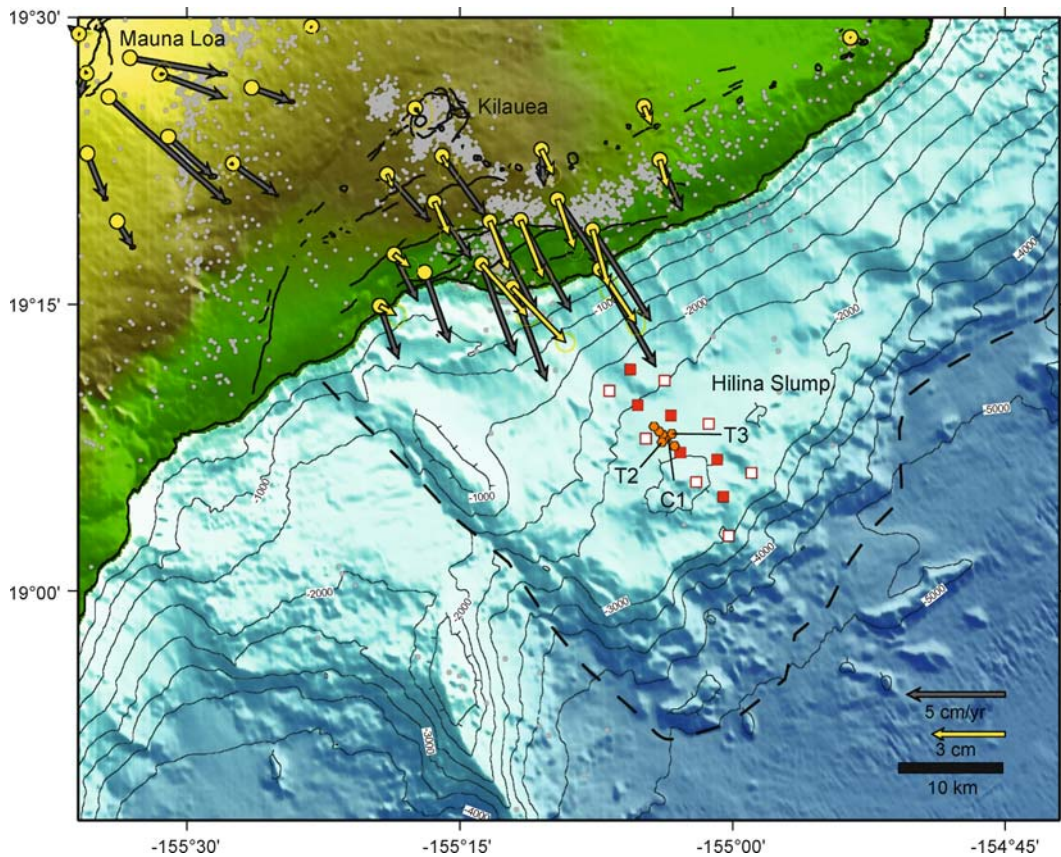
The morphology of submarine landslide features is similar to their subaerial counterparts. In map view they generally exhibit lobate and hummocky bathymetry. In cross-section, a wedge-shaped region of deformed material thins down-slope and is underlain by a gently-sloping planar dislocation surface (sometimes referred to as a basal detachment or 'decoulement') separating the deformed carapace from the underlying undeformed substratum (Fig. 2). An upslope extensional head-scarp region transitions into a contractional fold belt towards the toe. The normal faults in the upslope region typically intersect the surface at high angles ( $> 45$  degrees) and are manifested as scalloped-shaped scarps at the head of the slide that separate regions of differentially tilted fault blocks; in the sub-surface they may continue at high angles un-

til they intersect the basal decollement, or they may sole with depth either into the decollement or into another sub-horizontal slip-surface [24]. The contractional regions are characterized by folded and bulging layers, closed depressions, and steep toes [20,21,25]. Moore et al. [8] separated the Hawaiian submarine landslides into two principal types: 'slumps' and 'debris avalanches'. The slumps are wide (up to  $\sim 100$  km), deep-seated ( $\sim 10$  km thick), and have surface inclinations of up to 3 degrees while the debris avalanches are long (up to  $\sim 230$  km), shallowly-seated (50 m–2 km thick), and have surface slopes generally less than 3 degrees.

Concomitant with the mapping of the landslide features has been the increasing recognition that sudden submarine landslide movement can cause tsunamis with destructive implications for coastal societies e.g. [26]. A particularly well-known example of this scenario is the 1929 Grand Banks failure [27]. Moreover, in the 1990s alone workers have attributed at least 5 tsunami events to catastrophic landslide failure sources: (1) 1992 Flores Island, Indonesia [28]; (2) 1994 Mindoro, Philippines [29]; (3) 1998 Papua New Guinea e.g. [30]; (4) Kocaeli, Turkey [31]; (5) 1999 Pentecost Island, Vanuatu [32]. In Hawaii, the  $M_w$  7.7 1975 Kalapana earthquake, most likely



**Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 2**  
Schematic cross-section of a submarine landslide flanking an active ocean-island volcano (after [21])



**Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 3**

Map of geodetic networks on the Hilina Slump overlain on topographic/bathymetric map. Yellow circles on land are CGPS sites operated jointly by the USGS Hawaiian Volcano Observatory, University of Hawaii, and Stanford University. Red squares offshore (open and filled) are seafloor geodetic sites operated by Scripps Institution of Oceanography. Orange circles offshore are acoustic extensometer sites deployed by our group with locations of transponders (T2,T3) and transceiver (C1) indicated. Grey vectors are average horizontal velocities from 1997–2005. Yellow vectors are horizontal motions from the January 2005 slow earthquake. Grey dots are earthquakes from the HVO catalog for the period May 2004–2005

due to slip of the fault surface underlying an active submarine landslide [33,34,35], caused local loss of life and damage in Southern California.

From a hazards standpoint it is particularly important to understand how tsunami are generated by submarine

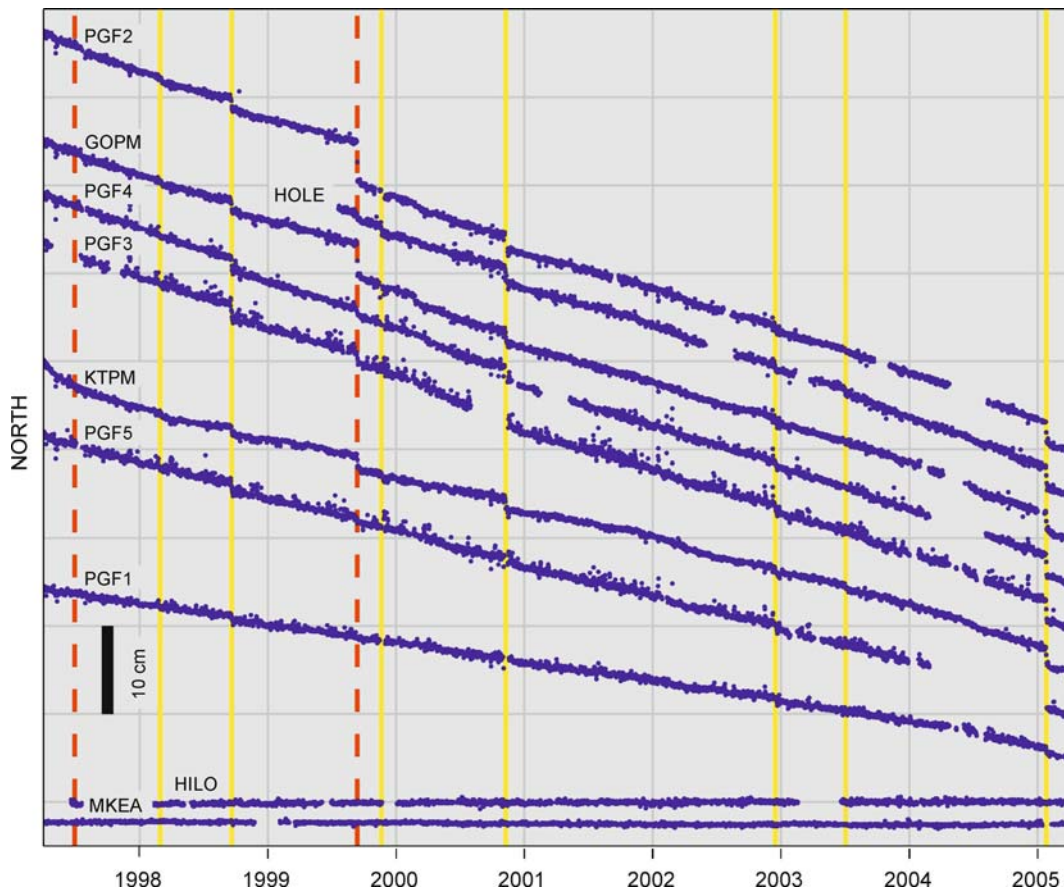
landslides because the propagation time between tsunami generation to runup is typically on the order of minutes. For instance, a catastrophic failure of the west side of the island of Hawaii would likely send tsunami waves around the Hawaiian islands that would reach the densely popu-



lated areas of Oahu's Waikiki beaches in less than an hour and more likely  $\sim 30$  min (G. Fryer, personal communication, 2007). Simulating waves generated by a sudden, chaotic disturbance of the seafloor, as expected from a submarine landslide, is quite complicated e. g. [36,37] and not all workers agree on approaches or results. Murty [38], however, stressed that parameters such as slide angle, water depth, density, speed, duration of the slide are second order, while instantaneously displaced volume is likely the most important parameter controlling tsunami generation.

Despite the increasing awareness of their hazard, little is known about how submarine landslides actually move, largely because of the challenge of installing instruments and retrieving data from the submarine environment. Moore et al. [8] recognized that while slumps more likely move relatively slowly, debris avalanches could be deposited very rapidly based on, for instance, uphill flow

of material in the distal portions of the landslide deposits. In agreement with this, the estimated downhill velocities from the 1929 Grand Banks event was 60–100 km/h [39] whereas many studies have documented  $\sim 6$ –10 cm/yr horizontal and vertical velocities associated with a submarine landslide flanking the Island of Hawaii's Kilauea volcano, the Hilina Slump [4,5,6,40,41,42]. Recently, GPS data from the Hilina Slump have elucidated that not only does the slump move at the above-stated, fairly smooth, background velocities, but also that the slump will occasionally deform in discrete accelerated cm-scale motions equivalent to  $M_6$  earthquakes but without the shaking [4,5,6]. These 'slow earthquakes' last hours and accommodate cm's of ground displacement (Fig. 3). It is not currently known, however, how the occurrence of a slow-earthquake in a submarine landslide system will affect its future movement by either making it more or less probable of failing catastrophically.



Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 4

Time series of north component of motion for selected GPS stations. Offsets of 7 slow earthquakes are identified by yellow lines. Red dashed lines are magmatic diking events

## Slow Earthquakes

The term ‘slow earthquake’ has been used to describe a variety of transient aseismic deformation phenomena including slow precursor events preceding large earthquakes [43,44,45,46,47,48], afterslip following earthquakes [49,50,51], variable fault creep rates [52,53,54,55], certain subduction zone thrust events with unusually long durations and large amplitude tsunamis for their size [56], and discrete fault-slip events that do not produce detectable seismic shaking but are accompanied by ground displacements very similar to those produced during earthquakes [1,2,3,4,5,6,57]. Hereafter, when we use ‘slow earthquake’ we will refer to this latter description although the term was first used in the modern literature to describe a slow precursor to the great Chilean 1960 earthquake [43].

In contrast to slow precursor events whose ground motions, like traditional earthquakes, are measured in seconds or minutes, slow earthquake (SE) displacements usually accrue over time periods ranging from hours to days and so they have been typically sensed with geodetic methods. In the last decade, the proliferation of continuous GPS (CGPS) networks has led to numerous SE observations and the discovery of some very rich behavior. In certain regions SEs have occurred with very regular periods [2,4,58], they are often associated with non-volcanic tremor [59,60] and, apparently, SEs follow very different scaling laws (moment vs. duration) than traditional earthquakes [61]. Explanations for SE slip behavior has varied to date. For subduction zones, the combination of deep (> 35 km) SE sources and their association with tremor (a phenomenon initially thought to be caused by forced fluid flow [62] but more recently also explained in terms of shear failure [63,64]) led to one current hypothesis that SE mechanics are controlled by water released during metamorphic phase changes at the interface between a subducting and overriding plate [65,66]. Other explanations have invoked rate- and state-variable frictional behavior to suggest that SEs occur preferentially at transitions between velocity strengthening and weakening regimes on a fault plane [67,68,69] and that temporally varying climatic load changes could help explain SE periodicity [58].

Due, in part, to their large magnitude deformation signal and the high concentration of CGPS networks focused on them, subduction zones have dominantly been the location of the most SEs to date [1,2,57,59,70,71,72]. Recently, the CGPS network on Kilauea volcano’s mobile south flank has recorded multiple SEs [4,5,6] (Fig. 4) and it is through the Kilauea events that SEs have come to be associated with submarine landslides.

## Monitoring Motion: Subaerial and Submarine Geodetic Methods

Not surprisingly, much more is known about the motion of subaerial than submarine landslides. Geodetic measurements on land combined with contemporaneous measurements of other properties (pore-water pressures, strength of materials, etc.) have allowed, in some cases, a very thorough understanding of how landslide motion is related to driving forces such as gravitational stresses and rainfall. For instance, Baum and Reid [73] instrumented a slow-moving submarine landslide in Honolulu’s Manoa valley with extensometers recording at 15 minute intervals and rain gauges and found a direct correlation between rain fall and deformation events. Similarly, Malet et al. [74] showed that GPS-measured surface velocities increased to as high as 20 cm/day following periods of higher rainfall during May 1999 at the Super-Sauze earthflow in the French Alps. At a slightly different scale, Hilley et al. [75] used InSAR to simultaneously map deformation of multiple landslides in California’s Berkely Hills at  $\sim$  monthly intervals and found that landslide motion correlated with times of high precipitation and that during the 1997–1998 El Nino event displacement rates doubled from the background rate of  $\sim$  27–38 mm/year, albeit with a  $\sim$  3 month time lag between the onset of motion and the high precipitation.

Much of the current knowledge that we have about the motion of submarine landslides comes from their easier-to-monitor subaerial portions, such as at Kilauea’s Hilina slump (Fig. 3). In the case of the Hilina Slump, fully 3/4 of the feature resides offshore at depths greater than 2000 m and the deformation monitoring network is necessarily concentrated on the down-dropped blocks near the head-wall scarp of the entire system. Recently, however, submarine methods have started to provide geodetic information from the ocean floor itself e.g. [7], e.g. [76,77,78,79]. As these methods become more cost-effective and widespread they will surely yield much insight into submarine landslide kinematics and, eventually, be employed in operational hazard monitoring/mitigation scenarios.

Whether a network is solely subaerial, submarine, or some combination of the two, it is important to consider that monitoring strategies can vary substantially depending on the time duration of the expected signal, the desired threshold of detection, and the desired time latency for individual solutions. For instance, simple detection and warning of catastrophic landslide failure needs the most rapid solution latency but coarse detection thresholds (meter rather than millimeter level, for instance) are appropriate. Conversely, if the goal is to detect small, potentially precursory motions such as slow earthquakes,



then detection threshold must be as sensitive as possible along with solution latency being low.

### Subaerial Geodesy: GPS

A multitude of geodetic techniques, from mechanically- to electromagnetically-based, are currently employed on landslides to measure motion caused by a range of deformational phenomena spanning opening of small surface cracks to the motion of kilometer sized blocks e.g. [74]. Data from Global Positioning System (GPS) networks has increasingly contributed to the library of observations associated with landslide motion. In particular, for submarine landslides that are large enough so that their motion causes earthquakes or slow earthquakes [4,6,80], a technique which is suitable for inter-station distances measured in kilometers is most appropriate and so below, we concentrate on the use of GPS with submarine landslides.

**GPS** GPS networks capable of sub-cm to mm-scale 3-dimensional ground motion detection are now deployed in many of Earth's most actively deforming zones (see for example, [http://sps.unavco.org/crustal\\_motion/dxdt/](http://sps.unavco.org/crustal_motion/dxdt/)) and readers are referred to thorough reviews of the general technique and its application for geodynamic studies [81,82,83].

Crustal motion GPS networks are usually divided into two types: those that record data continuously (CGPS) and those whose individual monuments are occupied less frequently in survey mode (SGPS). Depending on a variety of factors including the modernity of the receiver, the bandwidth of telemetry networks (should they exist), and the storage capacity of the archival center, CGPS sampling rate generally varies between once every 30 and 1 s (though most modern receivers are capable of sampling at frequencies higher than 1 Hz). SGPS sampling is more varied though it usually comprises re-occupation of sites at intervals ranging from months to years with occupation times of hours to days and sampling rates similar to CGPS. Accordingly, SGPS networks are more useful for wider ranging spatial characterization of deformation phenomena rather than for the rapid detection of motion or for tracking temporal evolution during transient events.

High rates of sampling alone, however, do not guarantee that CGPS network positional solutions will achieve their highest precision and/or accuracy. Assuming that all sites within a CGPS network have stable monuments and high-grade geodetic antennae and dual frequency receivers, the most important components of its error budget for deformation monitoring are: (1) integer ambiguity

resolution; (2) orbital estimation; (3) atmospheric delay estimation; (4) antenna multipath; (5) satellite constellation geometry, and (6) intra-network baseline length. For networks monitoring landslides with spatial scales on the order of kms or tens of kms, however, the baselines are short enough that errors scaling with baseline lengths are small contributors to the error budget. In addition, as absolute positioning in a global reference frame is not essential, precise orbital estimation is less important. For the other error sources, freely available software packages such as GAMIT, GIPSY, and BERNES (<http://facility.unavco.org/software/processing/processing.html>), and precise orbit processing centers such as the IGS (International GNSS Service, <http://igsceb.jpl.nasa.gov/>) or the Scripps Orbit and Permanent Array Center (SOPAC, <http://sopac.ucsd.edu/>) usually allow good enough mitigation of these errors so that daily GPS solutions have resolution on the order of 2–5 mm in the horizontal and  $\sim 2$ –3 times worse in the vertical e.g. [83]. This resolution rule-of-thumb, however, generally applies to post-processed data with occupation times exceeding  $\sim 6$ –8 h, use of precise orbits, and the best atmospheric estimation techniques e.g. [84]. Because of the hours-to-days delay needed for estimating various grades of precise orbits and the computational time needed to estimate all the parameters for all sites with this much data this standard rule-of-thumb cannot necessarily be applied to a real-time or near real-time solution.

**Real- and Near-Real Time GPS Processing** There are a variety of processing techniques that may be suitable for real-time or near-real time monitoring applications for the subaerial counterparts to submarine landslides. For instance, the well-known real time kinematic (RTK) positioning technique frequently employed by the surveying community applies differential corrections sent via radio link between stations to yield site position estimates with cm-scale precision over baselines up to  $\sim 10$  km in real-time [85]. The technique works best for baseline distances less than  $\sim 10$  km typically, because the assumption of correlated errors between stations is not necessarily valid and differential corrections cannot be accurately applied for larger baselines. Thus, RTK may be a suitable monitoring technique if a more spatially contained portion of a landslide, such as a fault zone, is a particularly good indicator of motion for the overall larger unit.

It is the number of measurement epochs required to resolve the integer-cycle phase ambiguity inherent with GPS data, however, that is the principal factor limiting the temporal latency of high resolution GPS positioning solutions. Kinematic techniques typically require minutes worth of GPS data (when sampled at 1–30 s) in order to re-

solve integer ambiguities for initialization and reinitialization if a cycle slip or loss of phase lock occurs during measurement, although progress is being made on mitigating ionospheric effects and reducing time of integer ambiguity resolution e. g. [86]. Regardless, if GPS station locations are chosen with relatively good sky-view then loss of lock due to poor satellite visibility should be minimal.

Recently, Bock et al. [87] developed a method of resolving integer ambiguities from a single epoch of dual frequency GPS phase and pseudo range data that provides independent epoch-by-epoch position estimates over baselines as large as  $\sim 40$  km. Their method allowed instantaneous positioning resolution of 1.5 cm in the horizontal and 7–8 times worse in the vertical coordinates. Langbein and Bock [88] applied the technique to determine offsets in positional time series due to slip events recorded by the Parkfield, California CGPS network which has similar spatial scales to a typical large submarine landslide. They found that offset sensitivity was  $\sim 5$  mm for a 2 s sampling window and this decreased to  $\sim 2$  mm when a 60 s window was used [88]. To the best of our knowledge this represents the current state-of-the art in terms of real-time detection of motion over spatial and temporal scales typical of submarine landslides.

At Kilauea, we have developed an automated *near* real-time processing strategy for our monitoring efforts at the Hilina Slump (Figs. 3 and 5) [89]. The CGPS data are telemetered hourly by radio modem back to the USGS Hawaii Volcano Observatory (HVO) where they are retrieved via FTP. We use a sliding window approach, collecting all data available within the most current two-hour period and performing a network solution using the PAGES [90] processing software. We use the 30 s, ionosphere free phase combination as the observable and use and hold fixed the IGS ultra-rapid orbits and apply the NGS antenna phase calibration patterns for each site [91]. We also apply the IERS standard solid Earth tide [92] and Schwiderski ocean tide loading corrections [93,94]. Every half hour we estimate a piece-wise, linear neutral atmosphere (troposphere) correction and one set of N–S and E–W neutral atmospheric gradient corrections for each site. We apply a ‘weak’ constraint of 10 cm to the atmospheric corrections based on examination of previous adjustments.

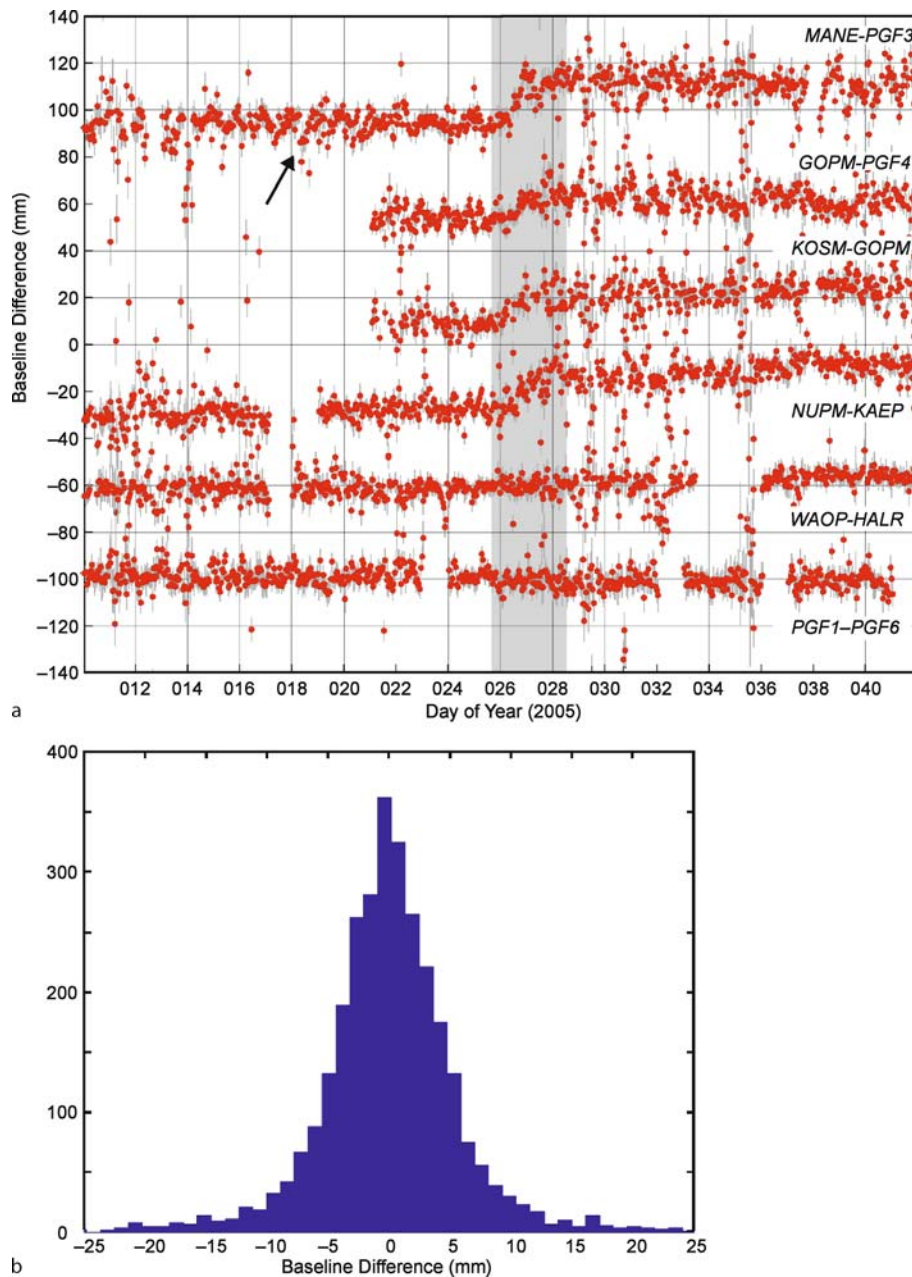
Figure 5a shows an example of baseline change time series from indicative stations derived from this approach and re-run in a simulated near real-time mode for the time period bracketing the slow earthquake from January 26–28, 2005 at Kilauea. Generally, the hourly baseline change noise levels are on the order of  $\pm 10$  mm, although because of some particularly large outliers, the standard deviation

of the baseline changes (not including the time period of the SE) is closer to 20 mm (Fig. 5b). Some of the large excursions in the time series not associated with the SE are due to the high amplitude, strongly spatially and temporally varying atmospheric water vapor gradients often associated with tropical islands such as Hawaii e. g. [95]. For instance for the excursion in the MANE-PGF3 baseline in the middle of day 18 (Fig. 5a), the time series takes a sharp upward bend until the start of day 19 when it again begins to oscillate about its mean value from present days. This baseline change gradient is very similar to the onset of the SE near the beginning of day 26. It is clear from the remainder of the time series after day 27 that the  $\sim 25$  mm offset is permanent, however, and so indicative of a real deformation event.

In a real-time monitoring scenario, because of the 2 h temporal latency and the atmospheric noise levels on the order of 1/3 the maximum signal levels in the baseline change plots, it would probably require on the order of 12–24 h before this event could be definitively classified as a deformational event. This could certainly be useful to hazard mitigators for being aware that a slow event such as an SE was occurring, but not in the event of a rapidly accelerating catastrophic collapse where the detection and warning time must be on the order of minutes. For more rapid warning, more sophisticated filtering techniques could be employed (such as monument motion, and network-wide coherence assessment as described below for the Network Inverse Filter [96]) although regions of large atmospheric gradient will always be hampered by low signal-to-noise ratios unless more sophisticated atmospheric mitigation techniques such as tomographic mapping [97] are employed in a real-time manner. Accordingly, the small detection threshold levels of the epoch-by-epoch processing at Parkfield described above [88] must be taken in the context of the relatively low atmospheric delay environment present there.

## Submarine Geodesy

Subaerial geodetic methods have, until now, provided essentially the entirety of geodetic evidence for submarine landslide-related deformation including slow earthquakes. Our understanding of submarine landslide motion and slow earthquake mechanics stands to increase dramatically in the coming decade, however, owing to the rapidly advancing field of submarine geodesy which can now reliably provide in situ measurements of seafloor deformation. The submarine environment is particularly challenging for geodesy techniques especially if they require the propagation of energy across a medium (ocean water) that



**Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 5**

**a** Baseline difference (position – median position) and  $2\sigma$  errors for near-real time processing of selected sites at the Hilina Slump during early 2005. See text for description of processing. Grey shaded region indicates the time interval of the slow earthquake. Black arrow indicates an example of an anomalous trend due to atmospheric delays discussed in text. **b** Histogram of baseline difference values for all of the sites in a, excluding the time period of the slow earthquake

can exhibit highly spatially and temporally variable material properties.

**Direct Path and Indirect Acoustic Approach** One of the submarine geodesy techniques most frequently em-

ployed to date is direct path acoustic measurement of baseline length changes using acoustic transponders. This method has been used to make in situ observations of tectonic motions on the seafloor at the Juan de Fuca ridge where one study found cm-scale motion occurring

over a period of a few days associated with an on-axis eruption [77] and another found no detectable motion over a period of a few years at a quiescent ridge segment [76]. By interrogating each other, pairs of instruments use two-way travel time measurements to constrain the inter-station distance. After correcting for variations in sound speed that result from changes in water temperature, salinity, pressure, and local tilt of the monument itself, these systems yield sub-cm precision for individual measurements. Operating frequencies typically range from 7.5 to 108 kHz and in deeper isothermal waters, the upward refraction of sound waves prevents signals from one transponder being received by the other at distances greater than  $\sim 1$  km [79]. Travel-time is determined simply by correlating the transmitted and received signals and picking the peak of the resultant correlogram. In quiet operating environments this can usually be done at the scale of  $\sim 5$   $\mu$ s, resulting in a  $\sim 4$  mm range error [7].

For the greater depth ranges, Sweeney et al. [79] devised an approach that allowed  $\pm 2$  cm resolution measurements for baselines up to 10 km at 2500–2600 m depths at a stable site on the Juan de Fuca plate. Sweeney et al. [79] suspended an acoustic interrogator hundreds of meters above the seafloor in a position acoustically visible to an array of seafloor-mounted instruments. In a manner similar to Spiess [98] they estimated the relative positions of the stations by moving the interrogator and collecting acoustic range data multiple locations.

The focus of these initial deployments of the direct and indirect acoustic approach systems was on measuring annual tectonic rates, and not necessarily on capturing transient events such as SEs that occur over hours or days. Accordingly sampling rates in current seafloor geodesy projects typically do not exceed a few times per day, largely because providing adequate power to seafloor instruments is still financially prohibitive. However with cabled oceanographic observatories scheduled to come online in 2007–2010 (<http://www.neptunecanada.ca/>; <http://www.orionprogram.org/>) the power delivery problem could be solved at high priority seafloor sites.

From October, 2005 through June, 2006 we deployed seven 10 kHz Linkquest transponders mounted  $\sim 3$  m above steel tripods and spaced over a distance of  $\sim 3$  km on the Hilina Slump (Fig. 3). One of the goals of this pilot project was to evaluate the performance of the transponders for use in submarine landslide monitoring at significant depth; each unit operated at a depth between 2640 and 2690 m. Battery power restrictions for the  $\sim 8$  month duration of the project meant that the instruments ranged to one another 12 times per day.

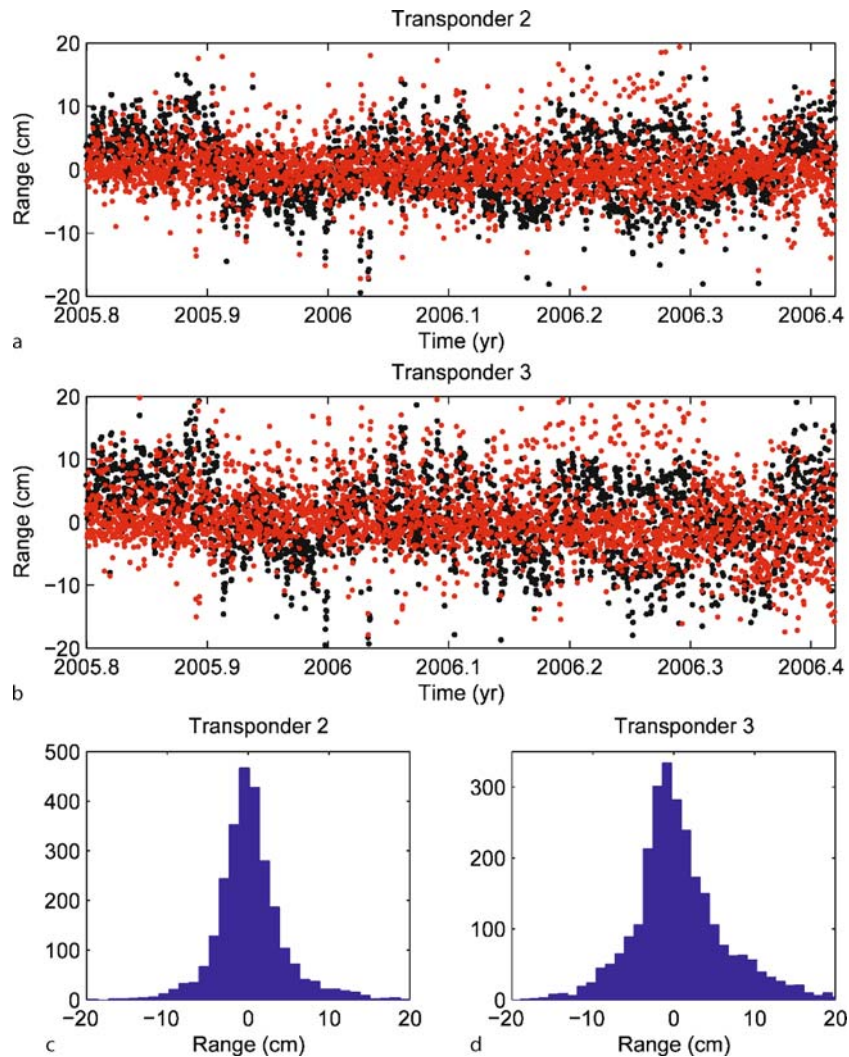
In Fig. 6 we show range change time series from two transponder-pair baselines ( $\sim 530$  and 683 m respectively) for which data recovery was complete over the experiment's duration. For the other transponder pairs data recovery was not as complete because either: (1) the unit was knocked over by local mass-wasting events or (2) the baseline distance was too long and data dropouts occurred when the transponders lost sync with one another. The two way travel times were picked from the peak of the correlation function between the outgoing and received waveforms at one end of the baseline. The black dots in Fig. 6a and b show the raw measurements, while the red dots show the distance measurements corrected for the variations in sound speed due to temperature changes, which were measured by an external conductivity and temperature sensor mounted on the transponder frames. The raw time series show significant long-term trends due to temperature variations that are effectively removed using just the temperature measurements at the transponders. We did not correct for salinity variations because all of our conductivity sensors provided contaminated data due to clogging by the local mass wasting events.

For these baseline pairs which were oriented approximately perpendicular to the maximum expected motion of the Hilina Slump we expect essentially no motion for such short baselines and during such a short measurement period. At the  $1\sigma$  level individual measurement noise is  $\sim 4.1$  and 5.7 cm respectively, though it is clear that smaller, cm-scale changes would be detectable given a long enough time period. For instance, daily estimates of baseline length have approximately a 1 cm standard deviation. For rapid event detection, baseline changes would need to be on the order of 10 or more cm or measurement frequency would need to be increased.

**GPS-Acoustic Method** The acoustic measurements described above provide only relative measurements of baseline length changes; however, for submarine landslide monitoring efforts it may be desirable to place submarine and subaerial measurements in a single regional or global reference frame. Largely with the aim of furthering seafloor active tectonics studies, a research group spearheaded by the late Fred Spiess conceived of, and have begun implementing integrated GPS and acoustic seafloor geodesy (GPS-A) studies, described in great detail by Spiess et al. [7] and references therein.

Briefly, GPS-A positions seafloor geodetic markers in a global reference frame by combining kinematic GPS positioning of a sea-surface platform (ship or buoy), precision underwater sound travel time measurement over km-scale path lengths, and a strategy for eliminating the large





**Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 6**

**a,b** Baseline difference (position – median position) for transponder-transceiver pairs (T2-C1, T3-C1 in Fig. 3) pairs from October 2005 – June 2006. See text for processing description. *Black dots*, raw baseline difference uncorrected for temperature. *Red dots*, baseline differences corrected for temperature. **c,d** Histogram of baseline difference values from **a** and **b**, respectively

errors in travel time measurement arising from a spatially and temporally varying sound speed structure in the near surface portion of the water column. The error mitigation strategy is based on the recognition that the location of the sea surface midpoint above the center of a triangle of seafloor transponders (themselves positioned via a near-bottom acoustic survey) is the point at which the three sea surface-to-seafloor travel times are identical and so, independent of sound speed.

Spiess et al. [7] reported repeatabilities of  $\pm 0.8$  cm and  $\pm 3.9$  cm in the north and east components, respectively, of seafloor measurements on the Juan de Fuca plate from

1994–1995. More recently, Gagnon et al. [78] reported velocities with  $\sim 5.5$  cm magnitudes and  $0.6$ – $0.8$  cm  $1\sigma$  errors from surveys between 2001–2003 on the updip portion of the Nazca–South America subduction zone in the Peru trench.

**Optical Path Length** Zumbege [99] combined commercial subaerial surveying technology (an EDM) with an optical fiber strainmeter in order to devise a low-power, cost-effective system for the harsh seafloor operating environment. EDMs (electronic distance meters) typically can measure distances with 1–2 mm precision over several km



by transmitting intensity modulated infra-red energy and measuring travel time from far away reflective surfaces. Zumberge [99] modified the EDM for seafloor geodesy by focusing the transmitted beam into the fiber's core and gathering the reflected light from the fiber's far-end into the EDM receiver optics.

The composite instrument comprises an optical fiber (enclosed in a hermetically sealed stainless steel casing) that stretches between two anchors, one 'active' and one 'passive', separated by several hundred meters on the seafloor. The active anchor houses the EDM and all supporting electronics and the passive anchor's purpose is to fix the fiber to the seafloor. In a test of their instrument, Zumberge et al. [100], report 1mm scatter of distance measurements over a 500-m-long fiber during a 50 day long period.

One of the obvious advantages of the optical path length technique is the high precision, isolation from sea water-borne errors, and the continuous nature of the measurement. The disadvantage, however, is that the technique is limited to short baselines ( $< \sim 1$  km) and so the precise location of straining regions must be known a priori. Nonetheless, for monitoring purposes with very well-defined targets such as the headwall portion of an active landslide, the technique holds much promise.

**Pressure Sensor Vertical Deformation** Phillips et al. [40] recently developed a new technique that used pressure sensors in campaign-mode repeat surveys to measure vertical deformation rates of the seafloor at depths exceeding 2000 m at the offshore portion of the Hilina Slump. Depth and pressure are related to one another in a straightforward manner through the hydrostatic equation, though for measuring seafloor deformation rates at the cm/yr level many other reductions must be performed (Phillips et al. [40] – also her thesis). The pressure sensor method is technologically challenging, requires much ship time, and requires during each site visit both a geodetic monument and a pressure sensor (lowered from a ship) to be operating in close proximity to one another at depth on the seafloor. Because, in each revisit it was not practical to collocate the pressure sensor in a repeatable fashion on the benchmark, Phillips et al. [40] found it necessary to measure the vertical offset between the pressure sensor and the benchmark via acoustic ranging.

Important error sources for this technique are poor knowledge of  $\alpha$  the specific volume of the water column (the reciprocal of density) and the speed of sound in depth. Due to changing tides, secular and seasonal barometric changes it is also necessary to utilize a common, stable ref-

erence, such as mean sea level (MSL) which, in turn, requires estimation of the geopotential anomaly.

Ultimately, Phillips et al. [40] conclude that their data show significant vertical deformation ( $9 \pm 2.4$  cm/yr) in the mid-section of the HS and negligible deformation towards the outer bench (just inboard of the toe of the entire landslide feature). These data are consistent with dislocation models delimiting the potential amount ( $25.0\text{--}28.1 \pm 7.3$  cm/yr), spatial extent ( $24.8\text{--}27.0 \pm 0.5$  km seaward of Kilauea's East rift zone), and depth (7 km) of the principal slipping surface below the landslide. This is a significant advance not only because it represents the first data set that allows a glimpse of how strain is partitioned in a massive submarine landslide such as the Hilina Slump, but also because it represents an ongoing monitoring project at a submarine landslide, albeit at temporally sparse sampling rate.

### Data Analysis and Inversion

In addition to the challenge of acquiring geodetic data from submarine or even subaerial landslides is the added challenge of inferring sub-surface fault geometry and/or deformation processes, with attendant realistic error bounds on estimated parameters, from a data set of earth surface displacements. We identify two principal types of analysis modes: (1) process-based and (2) hazards-based. Process-based analysis focuses on deriving the most accurate and, if possible, complete assessment of the factors involved in the observed landslide motion. Hazards-based analysis has two primary components: (a) rapid warning and (b) long-term hazards estimation. Rapid warning comprises, with the smallest temporal latency possible, providing information regarding the current state of the landslide system and how it relates to current, short- and long-term dangers to life and infrastructure. Long-term estimation comprises making probabilistic statements about the components and potential future behavior of the system and so temporal latency is not a limiting consideration.

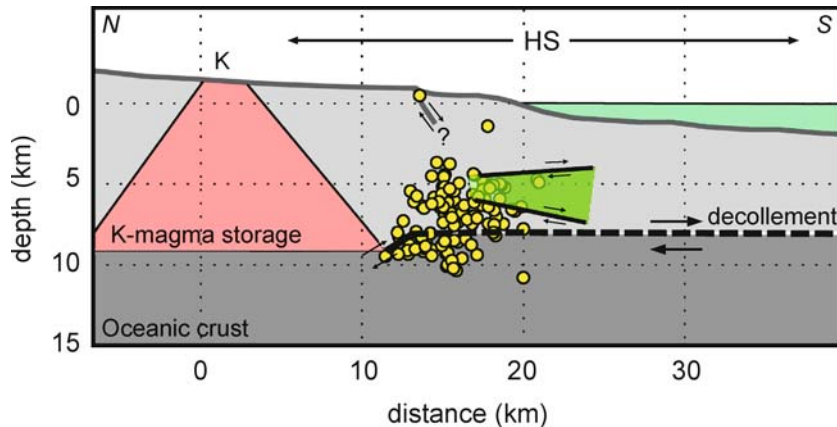
For either of the two types of analysis modes, the first objective is to usually try to relate the measured displacements  $d$ , of monuments at the surface of the earth to the source parameters,  $m$ , of the buried feature causing the deformation (usually considered a fault) through  $G(m)$ , a model of the deformation process. For instance, one very common model formalism, the dislocation in an elastic half-space, is a non-linear problem whose model vector  $m$  contains 9 parameters describing the location, orientation, and relative displacement of the dislocation approximating a faulting source [101,102].

Determining the most likely components of  $m$  and associated errors falls under the wide category of geophysical inversion for which there are a multitude of techniques e. g. [103]. In addition to the computational power available and the total amount of time allotted to the inversion, the choice of inversion method depends upon the nature of the model (for instance, linear vs. non-linear), the number of parameters in  $m$ , and the computational cost associated with individual model realizations and misfit assessments. Furthermore, it must be decided if the goal of an inversion is to obtain a rapid solution (a solution that satisfies some predetermined misfit level), an optimal solution (i. e. one 'best-fitting' solution), or rather to probe parameter space as thoroughly as possible with the goal of estimating posterior probability densities for each of the parameters for their value as indicators of resolution and uncertainty e. g. [104]. For instance, the most robust inversion method possible is the direct or 'grid' search, where the misfit for every possible permutation of  $m$  in parameter space is calculated. For more complicated or non-linear  $G(m)$  however the computational cost associated with each forward model run can make a grid-search time prohibitive, even for a process-based analysis. For these types of more difficult problems, inversion based on Monte Carlo sampling, which collects pseudo-random samples from multidimensional parameter space as a proxy for the problem's true posterior probability density,  $\sigma(m)$ , has been found to be quite successful [105,106]. Indeed, it is common practice in the tectonic geodesy community to use the Okada model combined with some type of Monte-Carlo method to derive parameters of earthquakes and slow earthquakes from surface displacements observed with GPS [80,104,107].

The recent evolution of the work on the SEs at Kilauea demonstrates how important the inversion results are to the overall analysis either from a process-based or long-term hazards-based perspective. In their initial recognition of the November 2000 SE, Cervelli et al. [5] used a simulated annealing optimization routine [80] and the Okada model to invert the GPS observations and conclude that the most likely dislocation source for the SE occurred on a gently landward-dipping thrust fault plane at 5–6 km depth. Combined with the fact that the SE post-dated by 9 days a burst of local rainfall of nearly 1 m and reasonable estimates of local hydrologic parameters, Cervelli et al. [5] suggested that the increased pore pressure due to deeply percolating rainwater triggered SE motion by inducing a  $\sim 2$  MPa pressure decrease of the effective normal stress on their preferred fault plane. Brooks et al. [4], however, showed from Gibbs Sampling inversion [104] results of GPS data on three additional SEs that posterior dis-

tributions of estimated fault parameters allow a wide family of equivalently plausible solutions, ranging from deeper seated decollement solutions at  $\sim 8$  km depth to the more shallow fault plane favored by Cervelli et al. [5] (Fig. 7). Moreover, they showed that the other SEs were not associated with anomalous rainfall. Segall et al. [6] then used the same inversion method of Cervelli et al. [5] for the GPS data of the additional SEs relocations of high-frequency earthquakes triggered by the January 2005 SE, seismicity rate theory [108], and Coulomb stress modeling [109] to conclude that the SE (and other similar events) likely occurred at a depth of  $\sim 8 \pm 1$  km on the main decollement plane below the Hilina Slump.

In this case, the crucial addition to the source inversion was the added constraint of the triggered microearthquakes which were relocated to depths near the decollement [6]. In their earthquake relocations, however, Segall et al. [6] did not use the full waveform data; rather, they used a double-difference-derived mapping with manual picks, assuming a 1D velocity model, between triggered events and previous high-precision relocations and tomography from elsewhere at Kilauea [110]. Two other studies, [111,112] performed high precision relocations using waveform cross correlation data and found different depths for the same cluster of events as Segall et al. [6]. Got and Okubo [111] suggested that their relocated events (including those triggered by the 1998 SE) do not illuminate a sub-horizontal fault plane but rather a deeper, steeply south-dipping reverse fault. Wolfe et al. [112] found that the triggered seismicity from the four SEs identified by Brooks et al. [4] consistently relocates on distinct map-view clusters aligned in the direction of the SE displacements themselves and in a subhorizontal band with depths of  $\sim 5$  km. This is a solution consistent with Morgan et al. [21] who used seismic reflection data to identify moderately landward-dipping fault planes at similar depths. While the epicenters are well constrained in the Wolfe et al. [112] study, however, they were concerned that poor station geometry as well as near source velocity heterogeneity may bias the absolute depth of the relocations and the analyses may not be capable of distinguishing between shallow and deep fault zones. Thus, while the decollement is certainly the prime candidate on which a SE would occur (in agreement with the Segall et al. [6] suggestion), the analyses remain somewhat equivocal. This, in turn affects our best understanding of the most important hazard issue, the propensity of the decollement for a catastrophic failure. In order to more definitively answer these questions the above-mentioned research groups in conjunction with the Hawaiian Volcano Observatory of the US Geological Survey (the group charged with hazards analysis and



**Submarine Landslides and Slow Earthquakes: Monitoring Motion with GPS and Seafloor Geodesy, Figure 7**

North-south cross-section of the Hilina Slump after Got and Okubo [111]. K, Kilauea. HS, Hilina slump. Black line below K is the seismic (solid) and aseismic (dashed) position of the decollement from Got and Okubo [111]. Two thicker black lines with green shaded area, are the family of geologically plausible dislocation solutions equivalently supported by inversion of GPS data. Yellow colored circles, locations of microearthquakes from HVO catalog for  $\pm 5$  days around each slow earthquake event

mitigation for the region) have teamed up in a joint seismological and geodetic research project aimed at definitively constraining the depths of triggered seismicity, and, hopefully, the depth of SE sources in the region.

While the above example focuses mostly on process and long-term hazards assessment, recent developments focused on deriving the time-dependent history of slip during a deformation event could also potentially be used for automated real-time inversion and event detection [96,113]. The Network Inverse Filter (NIF) [96] is a recursive Kalman filter algorithm that operates on either processed position estimates or raw phase data (rather than derived displacement rates) from an entire network of geodetic stations, includes a stochastic description of local benchmark motion, and finds a non-parametric description of slip rate on a fault plane as a function of time. In current implementations, the NIF employs Green's functions relating slip to surface displacement computed from the analytical solutions for a dislocation in an elastic half-space [101] although there is no reason other deformation models could not be employed [113]. Application of the NIF to the 1999 Cascadia SE allowed McGuire and Segall [113], for instance, to determine that slip rate on the slipping fault plane took up to  $\sim 20$  days to reach its peak and that the southern portion of the fault had finished slipping before the northern portion began to slip. Cervelli et al. [5] and Segall et al. [6] also applied the NIF to the Kilauea SEs to derive source-time functions for the events. One caveat, however, is that the fault plane is not solved for by the NIF, rather, it must be known and held fixed a priori. For subduction zones with greater spatial geodetic

coverage above the slipping portion of the fault this condition may be satisfied more satisfactorily than at submarine landslides.

### Discussion: Slow Earthquake and Submarine Landslide Process

Largely because of CGPS subaerial measurements at Kilauea there are now a suite of observations of slow earthquakes related to submarine landslide motion. As it seems that SEs are a fairly general fault slip phenomenon, occurring at a variety of subduction zone locales globally, it stands to reason that SEs may be common to submarine landslide-related deformation, at least on the flanks of ocean island volcanoes. Clearly, as seafloor geodesy projects become more common we will learn if this is the case or not. It is also clear that the search for a theoretical understanding of the slow earthquake phenomena will be coupled to further understanding the motion of submarine landslides.

Currently, explanations for slow earthquake slip behavior focus on rock mechanic theory that follows an empirically derived rate- and state-variable frictional constitutive law based on laboratory experiments e.g. [114]. (The description below follows the summary and notation in Scholz [114]). The Dieterich–Ruina or ‘slowness’ law is expressed as:

$$\tau = [\mu_0 + a \ln(V/V_0) + b \ln(V_0 \theta / \zeta)] \sigma,$$

where  $\tau$  is shear stress,  $\mu_0$  is the steady-state friction,  $\sigma$  is effective normal stress,  $V$  is slip velocity,  $V_0$  is a reference

velocity,  $a$  and  $b$  are frictional material properties,  $\zeta$  is the critical slip distance and  $\theta$  is a state variable that evolves according to:

$$d\theta/dt = 1 - \theta V/\zeta.$$

The frictional stability of the system, then, depends on  $(a - b)$ , the velocity dependence of steady-state friction, defined:

$$a - b = \delta\mu/\delta(\ln(V)).$$

In the context of a simple spring-slider model approximating a slipping fault, the boundary between the stable and unstable frictional regimes will occur at a critical value of effective normal stress,  $\sigma_c$ , given by:

$$\sigma_c = k\zeta / (a - b),$$

where  $k$  is the stiffness. When  $(a - b) > 0$  the material is said to follow ‘velocity strengthening’ behavior and the system is stable – earthquakes cannot nucleate in this regime and earthquakes propagating into such regions will be abruptly terminated. When  $(a - b) < 0$  the material is said to follow ‘velocity weakening behavior and the system is unstable for  $\sigma \geq \sigma_c$  – earthquakes will nucleate in this regime. When  $\sigma \leq \sigma_c$  the system exhibits oscillatory behavior and is said to be ‘conditionally stable’, it is stable under quasi-static loading but requires a discrete velocity perturbation in order for earthquakes to nucleate. Others interpret these relations slightly differently suggesting that nucleation occurs when matrix stiffness (scale and stress-rate dependent) drops to a critical value over a spatial scale large enough to promote rupture [68,115].

It may be in the boundary of this stability transition where slow earthquake slip behavior arises. For instance, a common observation of most of the subduction zone SEs is that they occur down-dip of the ‘locked’ zone or near the base of the seismogenic zone where earthquakes nucleate [1,57,58,69,71]. At these depths in subduction zones temperatures are close to the 450°C temperature at which feldspar starts to exhibit plastic behavior and so conditionally stable behavior would be expected [114], although it has recently been shown that transient oscillatory behavior may also arise naturally from system dynamics alone [116]. Faults in the conditionally stable regime, under steady-state loading, slip aseismically unless a perturbation to the system is large enough to push the fault across the stability boundary triggering an earthquake [117]. If the perturbation, however, is not quite large enough to push the fault across the stability regime then a period of stable sliding at increased velocity will occur

(i. e. a slow earthquake) as the fault evolves back to steady-state [117]. It is not yet clear, however, how the explanations for subduction-zone related SEs may translate to the submarine landslide environment. While many subduction zone SEs apparently occur down-dip of the ‘locked’ zone where earthquakes nucleate along the subduction interface, at Kilauea, SEs apparently occur up-dip of the zone where the majority of earthquakes occur [4,6].

Recent studies have more explicitly focused on helping to explain SEs in terms of the rate- and state-variable formalism. For instance, Kato [67] used rate- and state friction to simulate the effect that ‘asperities’, velocity weakening regions surrounded by velocity strengthening regions, have on slip behavior and found that episodic SEs occur when the velocity-weakening patch size is close to the critical size of earthquake nucleation. Similarly, Liu and Rice [68] found that when they applied along-strike variations in the frictional  $(a - b)$  parameter in models of subduction zone processes, that SE-type transient deformation events emerged spontaneously near the down-dip end of the seismogenic zone. They suggested that the down-dip end of the seismogenic zone is likely to be in the conditionally stable boundary between unstable- and stable regimes and that this could allow the SE behavior. More recently, Lowry [58] employed further theoretical implications of rate- and state-variable friction under resonant loading conditions [115,118] to suggest that Earth’s response to climatic redistribution of atmospheric, hydro-spheric, and cryospheric loads could lead to resonant fault slip behavior that could explain observed periodic slow earthquakes on the Cocos-North America plate boundary at Guerrero, Mexico.

Shibazaki and Shimamoto [119] recently proposed an alternate model of slow earthquakes that was motivated by laboratory experiments on the velocity dependence of frictional stability. They introduced a cutoff velocity to the rate-state formulation to mimic laboratory results where fault surfaces transition from velocity weakening at low sliding velocities to velocity strengthening at higher velocities. Implementing the laboratory values for these transitions reproduced the slip ( $10^{-7}$  m/s) and rupture propagation velocities (km/day) seen in subduction zone environments. Their model predicts a linear relationship between these two velocities that can in principle be inferred from high quality continuous geodetic measurements of slow earthquakes. This relationship may be an important way to relate observations of slow earthquakes to mechanical models, particularly in regions such as Kilauea where the fault planes are shallow enough ( $< 10$  km) to allow the details of the rupture to be resolved with high quality instrumentation. We note additionally that other recent

model parameterizations [116] yield recurrence intervals and propagation velocities similar to observed events and Lowry [58] also suggested means of relating geodetic observations and model parameters.

Two additional factors that may modulate the frictional properties of faults in submarine landslide environments include local hydrologic and magmatic forces (in regions of active volcanism) [23]. For instance, the Nov. 2000 SE at Kilauea post-dated by 9 days an intense rain storm of nearly 1 m at the southeastern Big Island [5]. Cervelli et al. [5] estimated permeability using regionally appropriate hydrologic parameters for porosity and fault zone diffusivity and suggested that the 1 m of rain could have triggered SE motion by inducing a  $\sim 2$  MPa pressure decrease of the effective normal stress on a gently landward dipping fault at  $\sim 5$  km depth. Brooks et al. [4], however, showed that 3 other SEs at Kilauea were not preceded by anomalous rainfall and that other periods of anomalous rainfall were not accompanied by SEs. Additionally, prior to these events Iverson [23] was skeptical of rainfall triggers as he found from a rigid wedge analysis that, in order for ground water head gradients to be large enough to destabilize the Hilina Slump, implausibly large clay layers ( $\sim 200$  m thick) or very low hydraulic diffusivity ( $\sim 10^{-11}$  m<sup>2</sup>/s) needed to be present. Others, however, have found that magmatic injection-induced mechanical and thermal pressurization of fluids may help to explain Canary and Cape Verde Islands flank instabilities [120]. Excess shear stresses exerted on a basal decollement because of rift zone magma injection have also been shown to be of sufficient magnitude to potentially cause slip [22,23]. Although others have suggested from geodetic observations of discrete dike events at Kilauea's rift zones that the dikes are injected passively, as a response to decollement slip, rather than as a trigger for it [121,122].

### Future Directions: Slow Earthquakes and Submarine Landslide Monitoring

It is clear that expanded seafloor geodetic monitoring of submarine landslides would be extremely important in bettering our understanding of both the slow earthquake process and the hazards associated with submarine landslides. For instance, at the Hilina Slump, it is not even known how much of submarine portion of the landslide actually displaces the seafloor during a slow earthquake. Given the logistical and financial challenges associated with seafloor geodesy, however, it is reasonable to ask if it is worth it to society to monitor submarine landslides. In the Hawaiian Islands, for instance, one compilation suggests that there have been at least 6 tsunami-generat-

ing landslide events in the past 300 000 years [123]. For comparison, the average 50 000 year recurrence interval for such events is 1 to 2 orders of magnitude larger than typical earthquake recurrence intervals in Southern California where substantial resources are focused on earthquake-cycle related monitoring (<http://www.wgcep.org/>). As discussed above, however, landslide-generated tsunamis can be quite damaging both locally and regionally. One study simulated  $\sim 30$  m wave heights reaching the California coast within 6 h of a massive collapse of the Hilina Slump at Kilauea [124]. Clearly minimizing the impact of such an event would be of societal benefit.

Continuous submarine landslide deformation monitoring, although expensive, is not out of the realm of possibility, as costs for seafloor geodetic instrumentation are decreasing. In the coming years, in order to make continuous monitoring efforts more feasible, research will likely focus on two major technical challenges: (1) instrument power delivery and (2) data transmission. While cabling a network via seafloor pathways is certainly one way of satisfying both requirements, cabling may not be the best long-term solution for a number of reasons. First, the costs of large cable lengths and laying them on the seafloor is often measured in the millions of dollars. Second, especially for submarine landslides, the seafloor across which the cable need be lain can be very rough and cable failure can be a quite common occurrence. Third, if cable runs are long then annual maintenance efforts and cost can also be quite high. Accordingly, techniques such as acoustic data transmission [125] and local power generation via a buoy, for instance, will be critical.

### Acknowledgments

We thank the Hawaiian Volcano Observatory of the US Geological Survey, in particular Asta Miklius, Maurice Sako, and Kevan Kamibayashi, for their collaboration and efforts in operating the CGPS network at Kilauea. We thank Paul Segall of Stanford University for sharing GPS data from this network. We thank Matt Gould and Nick Witzell of WHOI for engineering support. We thank Capt. Rick Myer and the crew of the R/V Kilo Moana for their help in deploying and retrieving the extensometers. We thank Mark Schenewerk for help with implementing PAGES. We thank Katie Phillips and Dave Chadwell for helpful discussion and sharing a manuscript in preparation. We thank Cecily Wolfe and Gerard Fryer for stimulating discussion. In particular we thank J. Moore, T. Lowry, and D. Chadwell for thorough and insightful reviews which improved this contribution. This research was supported by the Geophysics and Geochemistry pro-



grams of the US National Science Foundation, grants from School of Ocean and Earth Science and Technology at the University of Hawaii, and a grant from the Deep Ocean Exploration Institute of the Woods Hole Oceanographic Institution.

## Bibliography

1. Dragert H, Wang K, James TS (2001) A silent slip event on the deeper Cascadia subduction interface. *Science* 292:1525–1528
2. Miller MM, Melbourne TI, Johnson DJ, Sumner WQ (2002) Periodic slow earthquakes from the Cascadia subduction zone. *Science* 295:2423
3. Ozakawa S, Murakami M, Tada T (2001) Time-dependent inversion study of the slow thrust event in the Nankai trough subduction zone, southwestern Japan. *J Geophys Res* 106:782–802
4. Brooks BA, Foster JH, Bevis MF, Frazer LN, Wolfe CJ, Behn M (2006) Periodic slow earthquakes on the flank of Kilauea volcano, Hawai'i. *Earth and Planet Sci Lett* 246:207–216
5. Cervelli P, Segall P, Johnson K, Lisowski M, Miklius A (2002) Sudden aseismic fault slip on the south flank of Kilauea volcano. *Nature* 415:1014–1018
6. Segall P, Desmarais EK, Shelly D, Miklius A, Cervelli P (2006) Earthquakes Triggered by Silent Slip Events on Kilauea Volcano, Hawaii. *Nature* 442:71–74
7. Spiess FN et al (1998) Precise GPS/Acoustic positioning of seafloor reference points for tectonic studies. *Phys Earth Planet Inter* 108:101–112
8. Moore JG et al (1989) Prodigious Submarine Landslides on the Hawaiian Ridge. *J Geophys Res* 94:17465–17484
9. Krastel S et al (2001) Submarine landslides around the Canary Islands. *J Geophys Res* 106:3977–3997
10. Carracedo JC, Day SJ, Guillou H, Perez Torrado FJ (1999) Giant Quaternary landslides in the evolution of La Palma and El Hierro, Canary Islands. *J Volcan Geotherm Res* 94:169–190
11. Day SJ, Heleno da Silva SIN, Fonseca JFBD (1999) A past giant lateral collapse and present-day flank instability of Fogo, Cape Verde Islands. *J Volcan Geotherm Res* 94:191–218
12. Tucholke B (1992) Massive submarine rockslide in the rift-valley wall of the Mid-Atlantic Ridge. *Geology* 20:129–132
13. Driscoll N, Weissel JK, Goff JA (2000) Potential for large-scale submarine slope failure and tsunami generation along the US Mid-Atlantic coast. *Geology* 28:407–410
14. Moore JG (1964) Giant submarine landslides on the Hawaiian Ridge. *US Geol Survey Prof Paper D* 501:95–98
15. Moore JG, Fiske RS (1969) Volcanic substructure inferred from dredge samples and ocean-bottom photographs, Hawaii. *Geol Soc Am Bull* 80:1191–1202
16. Marti J, Hurlimann M, Ablay G, Gudmundsson A (1997) Vertical and lateral collapses on Tenerife (Canary Islands) and other volcanic ocean islands. *Geology* 25:879–882
17. Ward SN, Day S (2001) Cumbre Vieja Volcano – Potential collapse and tsunami at La Palma, Canary Islands. *Geophys Res Lett* 28:3397–3400
18. Clague DA, Moore JG (2002) The proximal part of the giant submarine Waialua landslide, Molokai, Hawaii. *J Volcan Geotherm Res* 113:259–287
19. Coombs ML, Clague DA, Moore GF, Cousens BL (2004) Growth and collapse of Waianae Volcano, Hawaii, as revealed by exploration of its submarine flanks. *Geochem Geophys Geosyst* 5: [doi:10.1029/2004GC000717](https://doi.org/10.1029/2004GC000717)
20. Morgan JK, Clague DA, Borchers DC, Davis AS, Milliken KL (2007) Mauna Loa's submarine western flank: Landsliding, deep volcanic spreading, and hydrothermal alteration. *Geochem Geophys Geosyst* 8:Q05002; [doi:10.1029/2006GC001420](https://doi.org/10.1029/2006GC001420)
21. Morgan JK, Moore GF, Clague DA (2003) Slope failure and volcanic spreading along the submarine south flank of Kilauea volcano, Hawaii. *J Geophys Res* 108:2415, [doi:10.1029/2003JB002411](https://doi.org/10.1029/2003JB002411)
22. Dieterich J (1988) Growth and Persistence of Hawaiian Volcanic Rift Zones. *J Geophys Res* 93:4258–4270
23. Iverson RM (1995) Can magma-injection and groundwater forces cause massive landslides on Hawaiian volcanoes? *J Volcan Geotherm Res* 66:295–308
24. Cannon EC, Bürgmann R, Owen SE (2001) Shallow normal faulting and block rotation associated with the 1975 Kalapana earthquake, Kilauea volcano, Hawaii. *Bull Seismol Soc Am* 91:1553–1562
25. Moore JG, Normark WR, Holcomb RT (1994) Giant Hawaiian landslides. *Ann Rev Earth Planet Sci* 22:119–144
26. Bardet J-P, Synolakis CE, Davies HL, Imamura F, Okal EA (2003) Landslide Tsunamis: Recent Findings and Research Directions. *Pure Appl Geophys* 160:1793–1809
27. Hasegawa HS, Kanamori H (1987) Source Mechanism of the Magnitude 7.2 Grand Banks Earthquake of November 18, 1929: Double-couple or Submarine Landslide? *Bull Seismol Soc Am* 77:1984–2004
28. Yeh H et al (1993) The Flores Island Tsunami. *Eos, Transactions, Am Geophys Union* 74:371–373
29. Imamura F, Synolakis CE, Titov V, Lee S (1995) Field Survey of the 1994 Mindoro Island, Philippines Tsunami. *Pure Appl Geophys* 144:875–890
30. Geist EL (2000) Origin of the 17 July 1998 Papua New Guinea Tsunami: Earthquake or Landslide? *Seismol Res Lett* 71:344–351
31. Yalciner AC et al (1999) Field Survey of the 1999 Izmit Tsunami and Modeling Effort of New Tsunami Generation Mechanism. *Eos, Transactions, Am Geophys Union* F751(abstract):80
32. Caminade JP et al (2001) Vanuatu Earthquake and Tsunami Caused Much Damage, Few Casualties. *Eos Trans Am Geophys Union* 81:641,646–647
33. Nettles M, Ekstrom G (2004) Long-Period Source Characteristics of the 1975 Kalapana, Hawaii, Earthquake. *Bull Seismol Soc Am* 94:422–429
34. Tilling RI et al (1976) Earthquake and Related Catastrophic Events, Island of Hawaii, November 29, 1975: A preliminary report. *US Geol Surv Circ* 740:33
35. Day SJ, Watts P, Grilli ST, Kirby JT (2004) Mechanical models of the 1975 Kalapana, Hawaii earthquake and tsunami. *Mar Geol* 215:59–92, [doi:10.1016/j.margeo.2004.11.008](https://doi.org/10.1016/j.margeo.2004.11.008)
36. Okal EA, Synolakis CE (2003) A Theoretical Comparison of Tsunamis from Dislocations and Landslides. *Pure Appl Geophys* 160:2177–2188
37. Ward S (2001) Landslide tsunami. *J Geophys Res* 106:11201–11215
38. Murty TS (2003) Tsunami wave height dependence on landslide volume. *Pure Appl Geophys* 160(10–11):2147–2153

39. Fine IV, Rabinovich AB, Bornhold BD, Thomson RE, Kulikov EA (2005) The Grand Banks landslide-generated tsunami of November 18, 1929, preliminary analysis and numerical modeling. *Mar Geol* 215:45–57
40. Phillips KA, Chadwell CD, Hildebrand JA (2008) Vertical deformation measurements on the submerged south flank of Kilauea volcano, Hawai'i reveal seafloor motion associated with volcanic collapse. *J Geophys Res* 113:B05106; doi:10.1029/2007JB005124
41. Owen S et al (2000) Rapid deformation of Kilauea Volcano: Global Positioning System measurements between 1990 and 1996. *J Geophys Res B: Solid Earth* 105:18983–18998
42. Delaney PT et al (1998) Volcanic spreading at Kilauea, 1976–1996. *J Geophys Res* 103:18003–18023
43. Kanamori H, Stewart GS (1979) A slow earthquake. *Phys Earth Planet Inter* 18:167–175
44. Sacks IS, Suyehiro S, Linde AT, Snoke JA (1978) Slow earthquakes and stress redistribution. *Nature* 275:599–602
45. Ihmlé PF, Jordan TH (1994) Teleseismic search for slow precursors to large earthquakes. *Science* 266:1547–1551
46. Kedar S, Watada S, Tanimoto T (1994) The 1989 Macquarie Ridge earthquake: Seismic moment estimation from long-period free oscillations. *J Geophys Res* 99:17893–17908
47. McGuire JJ, Ihmlé PF, Jordan TH (1996) Time-domain observations of a slow precursor to the 1994 Romanche transform earthquake. *Science* 274:82–85
48. Kanamori H, Cipar JJ (1974) Focal process of the great Chilean earthquake May 22:1960. *Phys Earth Planet Inter* 9:128–136
49. Burgmann R et al (2001) Rapid aseismic moment release following the 5 December, 1997 Kronotsky, Kamchatka, earthquake. *Geophys Res Lett* 28:1331–1334
50. Heki K, Miyazaki S-I, Tsuji H (1997) Silent fault slip following an interplate thrust earthquake at the Japan Trench. *Nature* 386:595–598
51. Segall P, Burgmann R, Matthews M (2000) Time-dependent triggered afterslip following the 1989 Loma Prieta earthquake. *J Geophys Res B: Solid Earth* 105:5615–5634
52. Linde AT, Gladwin MT, Johnston MJS, Gwyther RL, Bilham RG (1996) A slow earthquake sequence on the San Andreas Fault. *Nature (London)* 383:65–68
53. Gwyther RL, Gladwin MT, Mee M, Hart RHG (1996) Anomalous shear strain at Parkfield during 1993–94. *Geophys Res Letters* 23:2425–2428
54. Gao SS, Silver PG, Linde AT (2000) Analysis of deformation data at Parkfield, California: Detection of a long-term strain transient. *J Geophys Res* 105:2955–2967
55. Lienkaemper JL, Galehouse JS, Simpson RW (1997) Creep Response of the Hayward Fault to Stress Changes Caused by the Loma Prieta Earthquake. *Science* 276:2014–2016
56. Kanamori H, Kikuchi M (1993) The 1992 Nicaragua earthquake: a slow tsunami earthquake associated with subducted sediments. *Nature* 361:714–716
57. Lowry AR, Larson KM, Kostoglodov V, Bilham R (2001) Transient fault slip in Guerrero, southern Mexico. *Geophys Res Lett* 28:3753–3756
58. Lowry AR (2006) Resonant slow fault slip in subduction zones forced by climatic load stress. *Nature* 442(7104):802–805
59. Ito Y, Obara K, Shiomi K, Sekine S, Hirose H (2006) Slow earthquakes coincident with episodic tremors and slow slip events. *Science* 311:503–506
60. Rogers G, Dragert H (2003) Episodic Tremor and slip on the Cascadia Subduction Zone: The chatter of silent slip. *Science* 300:1942–1943
61. Ide S, Beroza GC, Shelly DR, Uchide T (2007) A scaling law for slow earthquakes. *Nature* 447:76–79; doi:10.1038/nature05780
62. Aki K, Fehler M, Das S (1977) Source mechanism of volcanic tremors: fluid driven crack models and their application to the 1963 Kilauea eruption. *J Volcan Geotherm Res* 141:259–287
63. Rubenstein JL et al (2007) Non-volcanic tremor driven by large transient shear stresses. *Nature* 448:579–582
64. Shelly DR, Beroza GC, Ide S (2007) Non-volcanic tremor and low-frequency earthquake swarms. *Nature* 446:305–307; doi:10.1038/nature05666
65. Peacock S, Wang K (1999) Seismic Consequences of Warm Versus Cool Subduction Metamorphism: Examples from Southwest and Northeast Japan. *Science* 286:937–939
66. Julian B (2002) Seismological detection of slab metamorphism. *Science* 296:1625–1626
67. Kato N (2004) Interaction of slip on asperities: Numerical simulation of seismic cycles on a two-dimensional planar fault with nonuniform frictional property. *J Geophys Res* 109:B12306; doi:10.1029/2004JB003001
68. Liu Y, Rice JR (2005) Aseismic slip transients emerge spontaneously in three-dimensional rate and state modeling of subduction earthquake sequences. *J Geophys Res* 110:B08307; doi:10.1029/2004JB003424
69. Miyazaki S, McGuire JJ, Segall P (2003) A transient subduction zone slip episode in southwest Japan observed by the nationwide GPS array. *J Geophys Res* 108(B2):2087; doi:10.1029/2001JB000456
70. Kostoglodov V et al (2003) A large silent earthquake in the Guerrero seismic gap, Mexico. *Geophys Res Lett* 30:1807; doi:10.1029/2003GL017219
71. Freymueller J et al (2001) The great Alaska 'earthquake' of 1998–2001. *EOS Trans Am Geophys Un* 82(47)
72. Hirose H, Hirahara K, Kimata F, Fujii F, Miyazaki S (1999) A slow thrust slip event following the two 1996 Hyuganada earthquakes beneath the Bungo Channel, southwest Japan. *Geophys Res Lett* 26:3237–3240
73. Baum RL, Reid ME (1995) Geology, hydrology, and mechanics of a slow-moving, clay-rich landslide, Honolulu, Hawaii. In: Haneberg WC, Anderson SA (eds) *Clay and Shale Slope Instability*. Geol. Soc. of Am, Boulder, Colorado
74. Malet JP, Maquaire O, Calais E (2002) The use of Global Positioning System techniques for the continuous monitoring of landslides: application to the Super-Sauze earthflow (Alpes-de-Haute-Provence, France). *Geomorphology* 43:33–54
75. Hilley GE et al (2004) Dynamics of slow-moving landslides from permanent scatterer analysis. *Science* 304(5679):1952–1955
76. Chadwell CD et al (1999) No spreading across the southern Juan de Fuca Ridge axial cleft during 1994–1996. *Geophys Res Lett* 26:2525–2528
77. Chadwick WW, Embley RW, Milburn HR, Meining C, Stapp M (1999) Evidence for deformation associated with the 1998 eruption of Axial Volcano, Juan de Fuca Ridge, from acoustic extensometer measurements. *Geophys Res Lett* 26:3441–3444

78. Gagnon K, Chadwell CD, Norabuena E (2005) Measuring the onset of locking in the Peru-Chile trench with GPS and acoustic measurements. *Nature* 434:205–208
79. Sweeney AD, Chadwell CD, Hildebrand JA, Spiess FN (2005) Centimeter-level positioning of seafloor acoustic transponders from a deeply-towed interrogator. *Mar Geol* 28:39–70, doi:10.1080/01490410590884502
80. Cervelli P, Murray MH, Segall P, Aoki Y, Kato T (2001) Estimating source parameters from deformation data, with an application to the March 1997 earthquake swarm off the Izu Peninsula, Japan *J Geophys Res* 106:11217–11237
81. Dixon TH (1991) An Introduction to the Global Positioning System and Some Geological Applications. *Rev Geophys* 29:249–276
82. Hager BH, King RW, Murray MH (1991) Measurement of Crustal Deformation Using the Global Positioning System. *Ann Rev Earth Planet Sci* 19:351–382
83. Segall P, Davis JL (1997) GPS applications for geodynamics and earthquake studies. *Ann Rev Earth Planet Sci* 25:301–336
84. Eckl MC, Snay RA, Soler T, Cline MW, Mader GL (2001) Accuracy of GPS-derived relative positions as a function of inter-station distance and observing-session duration. *J Geodesy* 75:633–640
85. Leick A (2004) GPS satellite surveying. Wiley, Hoboken
86. Warnant R, Kutiev I, Marinov P, Bavier M, Lejeune S (2007) Ionospheric and geomagnetic conditions during periods of degraded GPS position accuracy: 1. Monitoring variability in TEC which degrades the accuracy of Real-Time Kinematic GPS applications. *Adv Space Res* 39:875–880
87. Bock Y, Nikolaidis RM, de Jonge PJ, Bevis M (2000) Instantaneous geodetic positioning at medium distances with the Global Positioning System. *J Geophys Res* 105:28233–28253
88. Langbein, J (2004) Noise in two-color electronic distance meter measurements revisited. *J Geophys Res* 109:B04406; doi: 10.1029/2003JB002819
89. Brooks BA, Foster JF, Miklius A, Schenewerk M (2005) Extended GPS Network and Near Real-Time processing, Mauna Loa Volcano, Hawai'i. *Eos Trans AGU*, 86:Fall Meet Suppl
90. Schenewerk M, Dillinger W, Hilla S (2000) <http://www.ngs.noaa.gov/GRD/GPS/DOC/toc.html>
91. Mader G, MacKay JR (1996) Geoscience Laboratory, Office of Ocean and Earth Sciences NOS, NOAA. Silver Spring, Maryland
92. McCarthy D, Petit G (eds) (2003) IERS Technical Note 32
93. Cartwright DE, Edden AC (1973) Corrected tables of tidal harmonics. *Geophys J R Astron Soc* 33:253–264
94. Cartwright DE, Taylor RJ (1971) New computation in the tide-generating potential. *Geophys J R Astron Soc* 23:45–74
95. Foster J, Bevis M, Chen Y-L, Businger S, Zhang Y (2003) The Ka'u Storm (Nov 2000): Imaging precipitable water using GPS. *J Geophys Res* 108:4585
96. Segall P, Matthews M (1997) Time dependent inversion of geodetic data. *J Geophys Res* 102:22391–22409
97. Flores A, Ruffini G, Rius A (2000) 4D tropospheric tomography using GPS slant wet delays. *Ann Geophys* 18:223–234
98. Spiess FN (1985) Analysis of a possible sea floor strain measurement system. *Mar Geodesy* 9:385–398
99. Zumberge M (1997) Precise Optical Path Length Measurement Through an Optical Fiber: Application to Seafloor Strain Monitoring. *Ocean Eng* 24:532–542
100. Zumberge M et al (2006) In: Nadim F, Pottler R, Einstein H, Klapperich H, Kramer S (eds) 2006 ECI Conference on Geohazards. Lillehammer, Norway
101. Okada Y (1985) Surface deformation due to shear and tensile faults in a half-space. *Bull Seismol Soc Am* 75:1135–1154
102. Steketee JA (1958) Some Geophysical Applications of the Elasticity Theory of Dislocations. *Can J Phys* 36:1168–1198
103. Menke W (1984) Geophysical Data Analysis: Discrete Inverse Theory. Academic Press Inc, San Diego
104. Brooks BA, Frazer LN (2005) Importance reweighting reduces dependence on temperature in Gibbs samplers: an application to the coseismic geodetic inverse problem. *Geophys J Int* 161:12–21
105. Mosegaard K, Sambridge M (2002) Monte Carlo analysis of inverse problems. *Inverse Probl* 18:29–54
106. Kirkpatrick S, Gelatt CDJ, Vecchi MP (1983) Optimization by simulated annealing. *Science* 220:671–680
107. Hudnut KW et al (1996) Co-Seismic Displacements of the 1994 Northridge, California, Earthquake. *Bull Seismol Soc Am* 86:19–36
108. Dieterich J, Cayol V, Okubo PG (2000) The use of earthquake rate changes as a stress meter at Kilauea volcano. *Nature* 408:457–460
109. Stein RS (1999) The role of stress transfer in earthquake occurrence. *Nature* 402:605–609
110. Hansen S, Thurber CH, Mandernach MJ, Haslinger F, Doran C (2004) Seismic velocity and attenuation structure of the East Rift Zone and South Flank of Kilauea Volcano, Hawaii. *Bull Seismol Soc Am* 94:1430–1440
111. Got J-L, Okubo PG (2003) New insights into Kilauea's volcano dynamics brought by large-scale relative relocation of microearthquakes. *J Geophys Res* 108(B7):2337; doi:10.1029/2002JB002060
112. Wolfe CJ, Brooks BA, Foster JH, Okubo PG (2007) Microearthquake streaks and seismicity triggered by slow earthquakes on the mobile south flank of Kilauea Volcano, Hawai'i. *Geophys Res Lett* (In press)
113. McGuire JJ, Segall P (2003) Imaging of aseismic fault slip transients recorded by dense geodetic networks. *Geophys J Int* 155:778–788
114. Scholz CJ (1998) Earthquakes and friction laws. *Nature* 391:37–42
115. Perfettini H, Schmittbuhl J, Rice JR, Cocco M (2001) Frictional response induced by time-dependent fluctuations of the normal loading. *J Geophys Res* 106:435–438
116. Liu Y, Rice JR (2007) Spontaneous and triggered aseismic deformation transients in a subduction fault model. *J Geophys Res* 112:B09404; doi:10.1029/2007JB004930
117. Rice JR, Gu JC (1983) Earthquake aftereffects and triggered seismic phenomena. *Pure Appl Geophys* 121:187–219
118. Perfettini H, Schmittbuhl J (2001) Periodic loading on a creeping fault: Implications for tides. *Geophys Res Lett* 28:435–438
119. Shibazaki B, Shimamoto T (2007) Modelling of short-interval silent slip events in deeper subduction interfaces considering the frictional properties at the unstable-stable transition regime. *Geophys J Int* 171(1):191–205
120. Elsworth D, Day SJ (1999) Flank collapse triggered by intrusion: the Canarian and Cape Verde Archipelagoes. *J Volcan Geotherm Res* 94:323–340

121. Cervelli P et al (2002) The 12 September 1999 upper east rift zone dike intrusion at Kilauea Volcano, Hawaii. *J Geophys Res B: Solid Earth* 107:3-1–3-13
122. Owen S et al (2000) January 30, 1997 eruptive event on Kilauea Volcano, Hawaii, as monitored by continuous GPS. *Geophys Res Lett* 27:2757–2760
123. McMurtry GM, Watts P, Fryer GJ, Smith JR, Imamura F (2004) Giant landslides, mega-tsunamis, and paleo-sea level in the Hawaiian Islands. *Mar Geol* 203:219–233
124. Ward S (2002) Slip-sliding away. *Nature* 415:973–974
125. Frye D et al (2006) An Acoustically Linked Moored–Buoy Ocean Observatory. *EOS Trans AGU* 87:213–218

## Swarm Intelligence

GERARDO BENI

University of California Riverside, Riverside, USA

### Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Biological Systems  
 Robotic Systems  
 Artificial Life Systems  
 Definition of Swarm  
 Standard-Mathematics Methods  
 Swarm Optimization  
 Non-Linear Differential Equations Methods  
 Limitations of Standard-Mathematics Methods  
 Cellular-Computing Methods  
 Randomness in Swarm Intelligence  
 Asynchronous Swarms  
 The Realization of Asynchronous Swarms  
 Characteristics of Swarm Intelligence  
 Swarms of Intelligent Units  
 Future Directions  
 Bibliography

### Glossary

**Swarm Intelligence** (Definition 1, Sect. “[Definition of the Subject](#)”). The *intuitive* notion of “Swarm Intelligence” is that of a “*swarm*” of *agents* (*biological or artificial*) which, *without central control, collectively (and only collectively) carry out (unknowingly, and in a somewhat-random way) tasks normally requiring some form of “intelligence”*. (Definition 5, Sect. “[Swarms of Intelligent Units](#)”). The capability of universal computation carried out with natural asynchrony by a dynamic cellular computing system, none of whose cells can predict the computation done by the swarm.

**Swarm Robotics** The technology of robotic systems capable of Swarm Intelligence.

**Stigmergy** Indirect communication through modification of the environment.

**Pheromone** A chemical that triggers an innate behavioral response in another member of the same animal species

**Elementary Swarm** An ordered set of  $N$  units described by the  $N$  components  $v_i$  ( $i = 1, 2, \dots, N$ ) of a vector  $\mathbf{v}$ ; any unit  $i$  may update the vector, at any time  $t_i$ , using a function  $f$  of  $K_i$  vector components.

$$\forall i \in N: v_i(t+1) = f(\mathbf{v}_{k \in K(i)}(t))$$

**Cellular automaton** A system, evolving in discrete time steps, with four properties: a grid of cells, a set of possible states of the cells, a neighborhood, and a function which assigns a new state to a cell given the state of the cell and of its neighborhood.

**von Neumann architecture** [Computer](#) design that uses one [processing unit](#) and one [storage](#) unit holding both instructions and [data](#).

**Cellular-computing architecture** Computer design that uses cellular automata, and related machines, as processors and as storage of instruction and data.

**Intelligence** (working definition for Swarm Intelligence) Ability to carry out universal computation.

**Natural asynchrony** Asynchronous updating characterized by three properties: more than one unit may update at each time step; any unit may update more than once in each updating cycle; and the updating order varies randomly for every updating cycle.

**Optimization algorithms** Algorithms to satisfy a set of constraints and/or optimize (e.g., minimize) a function by systematically choosing the values of the variables from an allowed set.

**Swarm optimization** Ant colony optimization, Particle swarm optimization, and related probabilistic optimization algorithms.

**Ant colony optimization** Probabilistic optimization algorithm where a colony of artificial ants cooperate in finding solutions to optimization problems.

**Particle swarm optimization** Probabilistic optimization algorithm where a swarm of potential solutions (particles) cooperate in finding solutions to discrete optimization problems.

**Game of Life** A cellular automaton designed to simulate life-like phenomena.

**Dynamic cellular computing system** Cellular computing system whose cells are mobile.

**Unpredictable system** A system such that complete knowledge of its state and operation at any given time



is insufficient to compute the system's future state before the system reaches it.

### Definition of the Subject

The research area identified as "Swarm Intelligence" has evolved rapidly. The term "Swarm Intelligence" first appeared in 1989 [1,2]; by 2007 "Swarm Intelligence" was in the title of four books [3,4,5,6], two series of conference proceedings [7,8] and in a new technical journal [9], without mentioning other areas in which the term swarm itself has become popular. As the use of the term "Swarm Intelligence" has increased, its meaning has broadened to a point in which it is often understood to encompass almost any type of collective behavior. And since the term "Swarm Intelligence" has popular appeal, it is also sometimes used in contexts which have limited scientific or technological content. Other meanings, however, refer rigorously to precise concepts. The following treatment of Swarm Intelligence is based only on concepts that can be clearly defined and quantified. Hence it is more restricted than some broader Swarm Intelligence presentations but, even so, it describes an interrelated scientific/technical core which forms a solid basis for a well-defined multidisciplinary research area.

**Definition 1** The intuitive notion of "Swarm Intelligence" is that of a "swarm" of agents (biological or artificial) which, without central control, collectively (and only collectively) carry out (unknowingly, and in a somewhat-random way) tasks normally requiring some form of "intelligence".

A more specific definition requires a detailed discussion and so it is given at the end of the article. (Sects. "Characteristics of Swarm Intelligence" and "Swarms of Intelligent Units").

Although this notion of Swarm Intelligence might seem vague, we will see in the course of this article that in fact it has many specific implications. Note that the notion is broad, which partly explains its widespread use, but not so broad as to include any type of collective action of groups of simple entities, as will become clear later.

These characteristics of Swarm Intelligence are also those of several biological systems, e. g., some insect societies or some components of the immune system, so that Swarm Intelligence has become important for understanding certain mechanisms in biology.

Technologically, the importance of 'swarms' is mainly based on the potential advantages over centralized systems. The potential advantages are: (1) *economy*; the swarm components (units) are simple, hence, (in princi-

ple) mass producible, modularizable, interchangeable, and disposable; (2) *reliability*, due to the redundancy of the components; destruction/death of some units has negligible effect on the accomplishment of the task, as the swarm adapts to the loss of few units; (3) ability to perform *tasks beyond those of centralized systems*, e. g., escaping enemy detection.

From this initial perspective on potential advantages, the actual application of Swarm Intelligence has extended to many areas, described in the body of this article, and its potential for future applications remains high, as discussed in the concluding section.

Some current, proposed and/or potential applications are in defense and space technologies, (e. g., control of groups of unmanned vehicles in land, water, or air), flexible manufacturing systems, advanced computer technologies (biocomputing), medical technologies, and telecommunications.

### Introduction

Swarm Intelligence (SI) investigations were initially motivated by studies of groups of simple robotic units which offered promise for technology. These 'swarms' are modeled as collections of simple quasi-identical units with decentralized control and independent clocks [10]. The number of units is intermediate between those of typical systems investigated in physics and other traditional science fields; in fact, the swarm is of the order of  $10^2$  to  $10^5$  units, where  $s \ll 23$ , i. e., the swarm is not composed of so many units that statistical physics methods can be applied to it, nor of such a few units that its dynamic can be solved exactly (or numerically to high precision).

These features are typical also of many biological systems, such as insect societies. They are also the features of robotic systems potentially economic, reliable and capable of tasks beyond the capabilities of centralized systems.

There are various established fields of science and technology that deal to a certain extent with the intuitive notion of SI: Artificial Life, Ethology, Robotics, Artificial Intelligence, Computation, Self-Organization, Complexity, Economics, and Sociology. In all these fields there is at some level, or in some application, the need to understand, model, and predict the operation of groups of units which only by working together (in a not very structured way, and 'unaware' of the evolution of the group) carry out "intelligent" tasks. A simple illustration is provided by classic economics. In a free market economy, people trade with each other in an unstructured way: each makes independent decisions at unpredictable times and unaware of the global results. But the outcome is the solution to



a complex problem – the problem of correct pricing. Such a ‘swarm’ solves a problem not (or poorly) solvable by centralized control economies, thus exhibiting, in a certain sense, a high form of “intelligence”.

From this example it is clear that the intuitive notion of SI is easy to grasp. But it is not so easy to make it less vague and more precise and quantitative, i. e., to make it a useful working concept for science and technology. That the concept of SI is not easy to quantify follows from the difficulty of defining several of the key components of the intuitive notion of SI (Definition 1). First, “intelligence” is a notoriously ambiguous concept. Second, “randomly” is also not easily defined and quantified. Third, “only collectively” must be specified in terms of the critical number of agents required for the emergence of SI. In what sense is a unit ‘simple’ or, ‘un-intelligent’, and the task carried out ‘complex’ or, ‘intelligent’? Fourth, “unknowingly” implies that the global status and the goal of the swarm are, at least to some extent, unknown to the single agents. Which algorithms and communication schemes result in tasks carried out ‘unknowingly’ by the agents?

Because of these difficulties, in this article we first use the aforementioned (Definition 1) intuitive notion of SI to describe the current main areas of studies considered to be SI. This will provide an overview of the current status of the field; it will make it possible to quantify the four vague concepts (‘intelligence’, ‘randomly’, ‘collectively’, ‘unknowingly’) and to reach a more sharply defined concept of SI. From this, we will be able to see more clearly the limitations of SI, and so, its realistic potential for future applications.

In this article, the main areas of SI studies are described by making three very broad distinctions: (1) scientific interest vs. technological interest (Sects. “[Biological Systems](#)”–“[Definition of Swarm](#)”); (2) standard mathematics vs. cellular computational mathematics (Sects. “[Standard-Mathematics Methods](#)”–“[Cellular-Computing Methods](#)”); (3) synchronous operation vs. asynchronous operation (Sects. “[Randomness in Swarm Intelligence](#)”–“[Swarms of Intelligent Units](#)”). These distinctions in turn will provide a guide to clarifying the four vague concepts in the intuitive definition of SI (Definition 1) and, thus, a conceptual orientation for future studies and applications; they will also provide criteria for evaluating the promise of SI to solve complex problems that traditional approaches cannot.

Focusing on the first distinction (scientific vs. technological interest), the main scientific interest in SI originated with the work of biologists studying insect societies [3]. The main technological interest originated with roboticists trying to design distributed robotic sys-

tems [1,2]. A valuable reference on the development of SI is [3], dealing, in parallel, with these two interrelated interests.

## Biological Systems

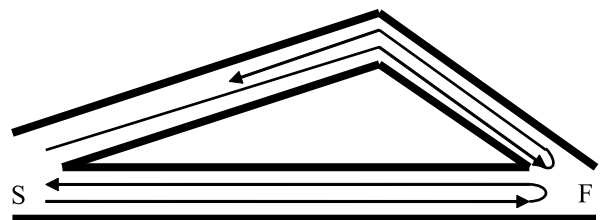
Probably the best known and seminal biology experiment in SI is the ‘double bridge’ experiment by Goss et al. [11]. While studying the foraging of ants they observed that if ants, starting from a point *S*, could reach food at a point *F* via two paths of different lengths, the ants would at first choose one of the two paths randomly; but after some ants had returned from *F* to *S*, more ants would choose to go from *S* to *F* via the shortest path; and eventually practically all the ants would choose the shortest path (see Fig. 1).

The key insight was the realization that the ants were finding the best path via *stigmergy*, that is, by *communicating through modification of the environment*. Ants are blind but they communicate chemically via pheromones. By laying pheromones along the path when returning from the food source *F*, the ants effectively marked the shortest path by laying more pheromones on it. After that, the ants that would start from *S*, would choose the path marked by more pheromones, i. e., the shortest path.

Thus it was observed and understood a method of self-organization and a method to solve a nontrivial problem by a form of collective intelligence, with many of the elements of the intuitive definition of SI given above.

Later, Dorigo [12] realized that this method could be abstracted and generalized to design algorithms that rely on ‘artificial ants’ to solve much more complex problems (see Sect. “[Swarm Optimization](#)”). Thus the close connection between biological studies of SI and its potential for technological application was first clearly demonstrated.

Many other experiments in ants and other social insects have confirmed the potential for developing bio-inspired algorithms [3,13,14,15]. For actual insect societies, various ant algorithms have been applied to model



**Swarm Intelligence, Figure 1**

Illustration of the double bridge experiments. The ants following the shorter path (the lower path) return to the source before the ants which have taken the longer path. In this way the shorter path has a higher density of pheromones; as a result, ants starting at *S* will now prefer the shorter path

tasks such as: division of labor, cemetery organization, brood care, carrying of large objects, constructing bridges, foraging, patrolling, chaining, sorting eggs, nest building and nest brooming. Social insects constitute 2% of insects, half being ants. Besides ants, termites, bees, and wasps, have been observed to exhibit some forms of SI behavior as in the aforementioned tasks.

Apart from insects, many other biological groups exhibit behavior with some of the features of SI, such as flocks of birds and schools of fish. A seminal model of artificial flocks and schools of fish was proposed by Craig Reynolds in 1987 [16].

It is a computational model for simulating the animation of a group of entities called “boids”, i. e., it is intended to represent the group movement of flocks of birds and fish schools. In this model, each boid makes its own decisions on its movement according to a small number of simple rules that react to the neighboring members in the flock and the environment it can sense. The simple local rules of each boid generate complex global behaviors of the entire flock. In addition to being used to simulate group motion in a number of movies and games, this flocking behavior has been used, e. g., for time-varying data visualization [17].

In studying these biological systems several concepts of relevance to SI were recognized. They can be summarized as:

- (1) Multiple communication (of various types) among units;
- (2) Randomness (random fluctuations);
- (3) Positive feedback, to reinforce random fluctuations;
- (4) Negative feedback for stabilization.

Of the various types of communication, we have already noted *stigmergy*, i. e., *indirect communication* by modification of the environment. On the other hand, *direct communication* may occur *unit-to-unit*, contact being a special case, (e. g., via antennae or mandibles in insects) or by *broadcasting* within a certain range (e. g., acoustically or chemically). The type and specific mode of communication has been found to be critical to the task performed, as e. g., in what types of patterns are formed [18].

Finally, the most basic lesson from biological studies of SI is that biology has found solutions to hard computational problems, and that the design principles used in doing this, can be imitated.

## Robotic Systems

The actual realization of SI systems as collections of robots is a very hard problem; in fact, it is quite difficult to make

even small groups of robots perform useful tasks [19,20]. Making even a single mobile, autonomous robot work in a reliable way (even in simplified environments) is a complex project. Often the technical problems with small groups of robots are quite far from the goal of SI, so there is not much reason to use the term ‘swarm’. Terms such as “collective robotics”, “multi-robot systems”, and “distributed autonomous robotic systems” are generally, and more appropriately, used. But, whenever the tasks carried out by these robotic systems become scalable to large numbers, the term “swarm robotics” is appropriate and, in fact, it has come into use. More typically, “swarm robotics” simply describes the design of groups of robotic units performing a collective task. Each robotic unit cannot solve the task alone; and collectively the robotic units try to accomplish a common task without centralized control.

As for any robotic system in general, each robotic unit, and the group as a whole, require design of: mechanics, control, and communications. The emphasis of current research, in relation to swarm robotics, is primarily on the latter two: (1) effective *communication* among the robot units [21], and (2) effective *control* via decentralized algorithms and robustness [22].

- (1) Research in robotic *communication* has become important with the growth of wireless communication networking and the lower cost of building robotic units, thus opening a new range of applications for multi-robot systems with networking capabilities, including swarm robotics. In fact swarm robotics provides the common ground for convergence of information processing, communication theory and control theory [21].
- (2) Research in *control* of robotic swarms is particularly important to guarantee the stability of the swarm since the swarm does not have a centralized control. The stability of a swarm is a special case of the general problem of distributed control. In fact, after swarm robotics algorithms for task implementation have been devised, the practical realization requires stability, and robustness, i. e., proper control. Swarm control presents new challenges to robotics and control engineers: various types of controllers for swarms are currently being investigated, e. g. neural controllers [23].

The control theory example coming closest to the problem of swarm control is perhaps that of ‘formation’ control, e. g., the control of multi-robot teams or autonomous aircrafts or land or water vehicles. These studies, when ex-

tended to decentralized systems, lead to consider problems of asynchronous stability of distributed robotic systems and swarms. [24]

Although much progress has been made in swarm robotics, the application of SI algorithms is still underdeveloped; one reason is that often the SI behavior emerges only above a critical number which is too large to make the construction of a robotic swarm practical, because too complex or expensive. Investigations of this type are thus generally carried out by simulation [22,25].

These simulations are specialized methods of swarm robotics. An example is 'executable models' [25] which can run in simulation or on a mobile robotic unit and can execute all aspects of a robotic unit behavior (sensing, information processing, actuation, motion), i. e., they fully represent how perception is translated into action by each robotic unit. They can be used to test either experimentally or in simulation how a group of robotic units behaves. Executable models are an evolution of early protocols (so called 'behavior-based' protocols) designed around the subsumption architecture [26]. Behavior-based protocols have now been generalized into Markov-type methods, i. e., protocols where the transitions between the possible states of a robotic unit are specified by a probability transition matrix as in Markov processes [27].

Looking at applications, swarm robotics has by now accumulated a collection of standard problems which recur often in the literature. One group of problems is based on *pattern formation*: aggregation, self-organization into a lattice, deployment of distributed antennas or distributed arrays of sensors, covering of areas, mapping of the environment, deployment of maps, creation of gradients etc. A second group of problems focuses on *some specific entity in the environment*: finding the source of a chemical plume, homing, goal searching, foraging, prey retrieval, etc. And a third group of problems deals with more *complex group behavior*: cooperative transport, mining (stick picking), shepherding, flocking, containment of oil spills, etc. This is not an exhaustive list: other generic robotic tasks, such as obstacle avoidance and all terrain navigation, are also swarm robotics tasks.

One envisioned application of swarm robotics which has received media attention is the ANTS (autonomic nanotechnology swarm) project by NASA [28,29]. This project envisions *nanobots* (i. e., a swarm of microscopic robots) operating autonomously to form structures for space exploration. The idea is inspired by the example of insect societies; it envisions a technology of self-similar, reconfigurable, miniaturized robotic units with a software strategy to endow the swarm with 'intelligence'. This ANTS 'intelligence' is at an intermediate level between tra-

ditional Artificial Intelligence (i. e., highly symbolic) and reactive responses 'intelligence', i. e., intelligence without internal representation [30]. Its basis is a software construct called a 'neural basis function' to bridge the gap between lower and higher level functions and to be capable of autonomous behavior. In one potential implementation, a Saturn autonomous ring array would launch 1,000 spacecraft with specialized instruments—organized as 10 subswarms—to perform in situ exploration of Saturn's rings to understand their constitution and formation.

The European Union sponsored swarm robotics project [31,32] was completed in 2005 after demonstrating several critical tasks, such as: autonomous self-assembly, cooperative obstacle avoidance, and group transport. For this project a new type of robot, called an s-bot was developed. A swarm-bot is any device composed of more than one s-bot, which is a mobile robot unit capable of connecting or disconnecting from another s-bot. S-bots have relatively simple sensors and motors and limited computational capabilities. Using their grippers, s-bots can assemble into a swarm-bot that is able to solve problems too difficult for a single s-bot. For example, a swarm-bot could transport an object too heavy for a single s-bot. [33]

Although swarm robotics could be defined as the robotic implementation of SI (Definition 1), so far, as noted, this implementation remains a distant goal. Meanwhile, concepts from SI can be usefully applied to collections of cooperating robots. Thus, referring to the intuitive notion of SI (Definition 1), the robotic swarm, can be characterized by the type of algorithm and of (decentralized) control, the number of units above which new behavior emerges, the communication method (range, topology, bandwidth), the processing and memory capability of each unit, and the heterogeneity of the group.

Swarm robotics, besides the implementation of SI algorithms, includes the material (mechanical and electronic) realization of the units comprising the swarm. This is, as noted, an arduous task which often becomes the emphasis of research in swarm robotics. But, as it was emphasized in the early years of SI, even if the material construction of the swarm were accomplished, SI algorithms would remain as the most difficult challenge for swarm robotics. This can be easily seen from the fact that a 'robot' swarm with very advanced hardware is already available for experimentation: it is a group of human beings. Each person could be limited in a controlled way, e. g., by allowing each person to handle only a specific device according to specific rules. Algorithms to make such a swarm doing intelligent tasks are in the province of SI, but they are not simple to devise, as common experience shows.

## Artificial Life Systems

The areas of *Self-organization*, *Complexity* and *Artificial Life* (or *A-life*) are all older and broader fields than SI and overlap with it to various extents.

A-life is conceptually placed somewhere between science and technology, and between biology and robotics. During the mid-1980s attempts at imitating living systems with machines grew rapidly and resulted in the formation of the research field of “Artificial Life” [34]. A-life investigates phenomena characteristic of living systems primarily through computational and (to a lesser extent) robotic methods.

Its scope is wide, ranging from investigations of how life-like properties develop from inorganic components to how cognitive processes emerge in natural or artificial systems. It includes research on any man-made systems that mimic the characteristics of natural living systems. By this criterion it includes SI, but actual, current A-life research is not much focused on SI; rather it focuses on origin and synthesis of life; evolutionary robotics; morphogenesis, learning, etc.

The basic theories at the foundation of A-life, and of relevance to SI, are the theories of *self-organization* and *complexity*. A-life studies systems which are typically characterized by many strongly coupled degrees of freedom. Systems of this type are more generally investigated within the science of *complexity* which began to be an active field of research in the early '80s. It is multidisciplinary and it investigates physical, biological, computational, and social science problems, including a vast range of topics [35] from environmental sciences to economics as it is clear from the content of this Encyclopedia. One basic feature that these systems have in common is the emergence of complex behavior from simple components, a notion we also find in SI.

In regard to self-organization, we note that, as many systems in nature, A-life systems may start disordered and featureless, and then spontaneously organize themselves to produce ordered structures, i. e., they self-organize. The theory of self-organization, going back to the 1950's [36], grew out of a variety of disciplines, but mainly from thermodynamics, non-linear dynamics and control theory. Self-organization can be defined as the spontaneous creation of a globally coherent (i. e., entropy lowering) pattern out of local interactions – a concept also relevant to SI.

Because of its distributed character, self-organization tends to be robust, resisting perturbations. The dynamics of a self-organizing system is typically non-linear, because of feedback relations between the components. Positive feedback leads to fast growth, which ends when all

components have been absorbed into the new configuration, leaving the system in a stable, negative feedback state. Non-linear systems have in general several stable states, and this number tends to increase (bifurcate) as an increasing input of energy forces the system away from its thermodynamic equilibrium. To adapt to a changing environment, the system needs a variety of stable states that is large enough to react to perturbations but not so large as to make its evolution uncontrollably chaotic. The most adequate states are selected according to their fitness, either directly by the environment, or by subsystems that have adapted to the environment at an earlier stage.

Formally, the basic mechanism underlying self-organization is the (often driven by randomness) variation which explores different regions in the system's state space until it enters an attractor. This precludes further variation outside the attractor, and thus restricts the freedom of the system's components to behave independently. It is equivalent to the decrease of statistical entropy that defines self-organization.

It is useful to keep this brief sketch of self-organization theory in mind as we proceed in describing SI, since the concepts in the theory of SI are evolved from a combination of concepts of self-organization and computation.

## Definition of Swarm

After having looked, in the previous three sections, at actual robotic, biological, and A-life systems and ideas of complexity and self-organization related to SI, we can return to the intuitive definition of SI (Definition 1) and make it more quantitative.

The intuitive notion consists of four elements: SI is “intelligence” achieved “collectively”, “randomly”, and “unknowingly”. An elementary swarm retaining these four elements can be defined as

**Definition 2 Elementary Swarm** An ordered set of  $N$  units described by the  $N$  components  $\mathbf{v}_i$  ( $i = 1, 2, \dots, N$ ) of a vector  $\mathbf{v}$ ; any unit  $i$  may update the vector, at any time  $t_i$ , using a function  $f$  of  $K_i$  vector components.  $\forall i \in N: \mathbf{v}_i(t+1) = f(\mathbf{v}_{k \in K(i)}(t))$

The Elementary Swarm describes an internally driven “collective” action. External input may be added in  $f$ . “Randomness” is built in the updating times. The evolution occurs “unknowingly” since the units have no processing capability. The Elementary Swarm can be generalized so that randomness appears also in the parameters of the function  $f$ . A further generalization is obtained by letting each vector component to be not just one number but a set of parameters.

Hereinafter we call ‘*Swarm*’ (capital S) any system capable of SI. It is worth noting that even in Swarms more general than the Elementary Swarm, the modeling is assumed restricted in such a way that *no unit is capable of computing the Swarm’s next global state*, (see also Sect. “[Swarms of Intelligent Units](#)”). Finally, “intelligence” is expected to be achieved by running appropriate algorithms via the updating function  $f$ . If and how this is going to be possible requires a more mathematical discussion, which is the subject of Sects. “[Standard-Mathematics Methods](#)” – “[Cellular-Computing Methods](#)”.

### Standard-Mathematics Methods

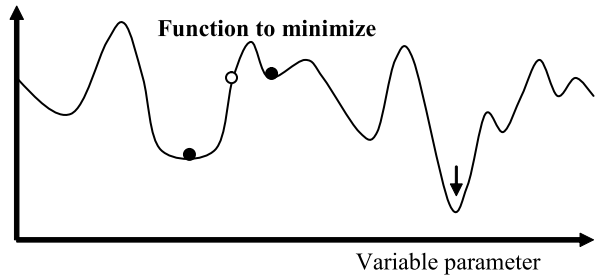
The science of biological swarms and the engineering of robotic swarms, as well as research in A-life relevant to SI, have progressed by using a broad range of mathematical techniques. All these techniques can be classified in two main groups: (1) ‘*standard-mathematics*’ methods and (2) *cellular-computational methods*.

By *standard-mathematics methods* (SMm) we mean any method that is based on the standard tools of applied mathematics and computations based on standard (Von Neumann) computer architectures. Examples are methods in differential equations, stochastic techniques, linear systems, and optimization. By *cellular-computing methods* (CCm) we mean highly parallel and local computational methods, with simple cells as the basic units of computation, typically carried out on ► [Mathematical Basis of Cellular Automata, Introduction to](#) (CA) [37].

These two mathematical approaches reflect two distinct trends in the evolution of SI research, as described below. We consider first (Sects. “[Swarm Optimization](#)” – “[Limitations of Standard-Mathematics Methods](#)”) the approach to SI based on SMm, since the greatest number of significant results in the area of SI has been obtained, so far, by standard-mathematics methods; specifically in the areas of *optimization* and *non-linear dynamics*. We consider them in turn in the next two sections.

### Swarm Optimization

Optimization is by far the largest research area associated with SI. This is due to two extremely successful optimization methods, whose origin is related to models of SI. The two methods are the Ant Colony Optimization (ACO) [12,13] and the Particle Swarm Optimization (PSO) [4,38]. Both ACO and PSO, originated in the early nineties and have resulted in hundreds of applications based on variations of the original algorithms. So much so that the field of “Swarm Optimization” could stand alone, apart from its relation to SI with which it is some-



**Swarm Intelligence, Figure 2**

Simplified illustration of the typical problem encountered in optimization. Starting from the value represented by the open circle and varying the parameter continuously the algorithm will find one of the nearest local minima (black circles) rather than the global minimum (indicated by the arrow)

times even identified. A thorough and recent description of swarm optimization techniques is in [6]. Here only the key concepts of swarm optimization are reviewed, as they relate to SI.

In PSO and ACO, as in any optimization methods, a function must be optimized, e.g., minimized. To find the minimum of the function, the variable is changed in a systematic way – the optimization method. Generally, the variable spans a multidimensional space. The search for the global minimum is nontrivial since the function may have many local minima, and the search could end into one of them (see Fig. 2). Various techniques to avoid this trapping have been developed by using some degree of randomness in the search strategy. For example, simulated annealing [39] was developed to overcome the limitations of non-random methods, e.g., the gradient descent [40]. PSO and ACO belong to this class of optimization techniques that make use of randomized searches.

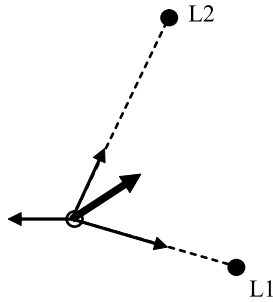
### Particle Swarm Optimization (PSO)

PSO was developed by Kennedy and Eberhart in 1995 [38] inspired by the social behavior of bird flocking and fish schooling [16].

In PSO, the position of each unit of the swarm is a point in the variable space of the function to be minimized. Every unit tries to reach the position corresponding to the minimum of the function. Each unit is assumed to know the global minimum value of the function, and to detect the value of the function at its location as well as the value of the function at the locations of a group of neighbors. The size of the group of neighbors is a parameter and could be the whole swarm.

The algorithm is, schematically, as follows. The units are initially in random locations. Every unit moves in





**Swarm Intelligence, Figure 3**

Illustration of the velocity update mechanism in PSO. A unit (white circle) originally moving to the left, changes its velocity by adding the two components in the direction of L1 (the unit's best location so far) and of L2 (the neighbors' best location so far). The magnitude of these components is determined by random weights. The new velocity (heavy solid line) generally tends to be in the direction of the global optimum

the variable space and remembers the location L1 where, among the locations visited so far, the function had minimum value. It also remembers the location L2 where, among the locations visited by all its neighbors, the function had minimum value.

At each time step, each unit calculates its distances from L1 and from L2; forms a weighted average of L1 and L2 using random weights; and changes its velocity in proportion to this weighted average (see Fig. 3). As a result every unit tends to move toward the location of the global minimum by taking advantage of its and its neighbors' knowledge. The process stops either when a unit is sufficiently close (by a chosen tolerance) to the location of the global minimum or when a chosen maximum number of iterations has been run.

PSO belongs to the category of stochastic, population-based algorithms, such as, e.g., genetic algorithms. The PSO's great merit is its simplicity. In many cases it outperforms genetic algorithms. Similarly to all optimization algorithms of this type, PSO convergence relies on the use of heuristics; and convergence does not mean convergence to the optimum. In fact, the basic PSO does not guarantee convergence even to a local minimum. The basic PSO is also inefficient in dynamic optimization problems, i.e., problems in which the optimum location changes. However, variations of the basic PSO have been proven to have improved performance in dynamic problems and to be capable of convergence to local minima.

Many improvements of the PSO basic algorithm have been developed and applied successfully to a wide range of optimization problems: continuous and discrete, constrained and unconstrained, single and multi-objective, static and dynamic. Specific applications cover just about

all areas of applied optimization. The main classes of applications have been in areas such as:

- (1) Neural networks (training, supervised and unsupervised learning, architecture selection, etc.);
- (2) Game learning;
- (3) Clustering;
- (4) Design (aircraft wings, antennas, circuits);
- (5) Scheduling & Planning (maintenance, traveling salesman, power transmission, etc.);
- (6) Controllers (flight path, air temperature, power stabilizers, etc.);
- (7) Data mining. For more detail see, e.g., [6].

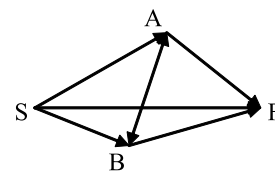
### Ant Colony Optimization (ACO)

The key idea of ACO [12,13] is an abstraction and generalization of the Goss et al. [11] two-path experiment with ants. A first generalization of the two-path problem is finding the shortest path between the starting point S and the final point F when between S and F there are many possible paths.

The two-path problem can be represented by a graph with 3 vertices (S, A, F) and three arcs ( $S \rightarrow A$ ,  $A \rightarrow F$ ,  $S \rightarrow F$ ). The short path is  $S \rightarrow F$  and the long path is  $S \rightarrow A \rightarrow F$ .

If we add another vertex B, and make the graph complete (i.e. we join with an arc each vertex to every other vertex) we obtain five possible paths:  $S \rightarrow F$ ,  $S \rightarrow A \rightarrow F$ ,  $S \rightarrow B \rightarrow F$ ,  $S \rightarrow A \rightarrow B \rightarrow F$ ,  $S \rightarrow B \rightarrow A \rightarrow F$  (see Fig. 4). The idea is easily generalized to a complete graph with more vertices. The number of paths increases exponentially so checking the length of all the possible paths becomes computationally unfeasible for a large enough number of vertices.

A number  $N$  of 'ants' start at vertex S choosing randomly which vertex to go next. At every new iteration, each ant decides which arc to traverse next, with probability proportional to the amount of pheromone on the arc relative to the total amount of pheromone on the possible arcs that the ants could choose.



**Swarm Intelligence, Figure 4**

The five possible paths from S to F:  $S \rightarrow F$ ;  $S \rightarrow A \rightarrow F$ ;  $S \rightarrow B \rightarrow F$ ;  $S \rightarrow A \rightarrow B \rightarrow F$ ;  $S \rightarrow B \rightarrow A \rightarrow F$

After an ant,  $k$ , reaches the destination  $F$ , the length of its path  $L_k$  is remembered (if an ant reaches the destination via loops,  $L_k$  is calculated after removing the loops). The ant retraces exactly the path (without loops) and deposits pheromones on the arcs in proportion to  $1/L_k$ .

In this way, the marking (with pheromones) of the paths by all the ants modifies the graph so that the probability of any ant taking the shortest path at the next trip from  $S$  to  $F$  increases. Eventually all the ants will follow the same path, if the algorithm converges, and the path will be the shortest if the convergence is to the global minimum. As in most optimization methods, this is never guaranteed.

The idea of the original ACO algorithm [12] adds to the foregoing sketch of the basic model three more elements. First, each arc has an a priori propensity to be traversed (regardless of pheromone content); second, each ant keeps in memory a tabu list of arcs not to be traversed, to avoid loops; and third, the pheromones evaporate at a given rate.

ACO key insight is the application of the concept of stigmergy to stochastic optimization. The ants communicate by modifying the environments (graph) and act probabilistically on the basis of the modified environment.

Many variations of the basic ACO algorithm have been proposed and implemented. The many variations take advantage of specific knowledge about the specific problem, i. e., they use heuristics, e. g., by setting the a priori propensity of traversing an arc, or by setting the evaporation rate.

ACO eventually resulted in a meta-heuristic which is a strategy for designing ACO heuristics. Various ACO-based metaheuristics have been developed. Similarly to PSO, ACO algorithms have been applied to all the basic types of optimization problems: continuous and discrete, constrained and unconstrained, single and multi-objective, static and dynamic. The first application of ACO was to the Traveling Salesman Problem, which is an NP-hard combinatorial optimization problem, and it is the most frequently attacked problem using various ACO heuristics. The main classes of other applications are to problems of

- (1) Ordering (scheduling, routing);
- (2) Assignment (Neural network training, image segmentation, design);
- (3) Subsets finding (maximum independent set);
- (4) Grouping (clustering, bin packing).

Clearly “swarm optimization” successfully uses concepts from the general notion of SI, but optimization is not in itself a necessary characteristic of SI. In fact, many tasks actually or potentially carried out by swarms are not-optimal in any sense.

## Non-Linear Differential Equations Methods

One fruitful approach to modeling swarms has been to treat each individual as a discrete particle. These “individual-based” models have been employed in quite a few biological and mathematical studies. They are based on simple rules of motion for each individual, involving some combination of self-propulsion, random movement, and interaction with neighboring organisms. The models typically take the form of coupled **non-linear** difference or differential equations, which may be stochastic or deterministic, depending on the particular features of each model. Numerical simulations have revealed collective behavior. But a main disadvantage of such models is that, for realistic numbers of individuals, analytical results for the collective motion are difficult or impossible to obtain. It is worth mentioning that some progress has been made in obtaining analytical results for stationary groups. In [41], a discrete model was formulated, and a Lyapunov functional was used to successfully predict an equilibrium state of equally spaced organisms. However, analytical (nonstatistical) descriptions of nonequilibrium states in discrete swarm models are few.

Other investigations of swarming have been carried out in a continuum setting, in which relevant quantities are described as scalar or vector fields. This approach goes back to 1980; reviews are provided in [42]. Continuum models may be constructed a priori or by coarse-graining a particle model. In general, continuum models provide a convenient setting in which to study large populations, since one may apply machinery from the analysis of *partial differential equations*. In the context of swarms, the focus has generally been on models in which the population density satisfies a convection-diffusion equation ensuring that the population density is conserved while individuals travel with a set average velocity. Recent models of this type [43] can predict, e. g., whether a population aggregates or disperses, the regions of aggregation, and length scales of the density patterns.

## Limitations of Standard-Mathematics Methods

In describing the SI investigations in the previous two sections, we have encountered the concepts of *optimization* and *non-linearity* (and earlier, in Sect. “*Artificial Life Systems*”, *complexity* and *self-organization* arising from non-linearity). To see more precisely the relation of these four concepts to SI we refer back to the definition of Elementary Swarm (Definition 2). Note that, by this definition, the Swarm is in principle capable of optimization, complexity and self-organization. In fact, it is clear that a Swarm is a *self-organizing* system, by definition, and that, depend-

ing on the choice of the updating function  $f$ , the pattern formed by the swarm components might, in principle, achieve high *complexity*.

As for non-linearity, it is convenient to think of  $f$  as a function of a function,

$$f = f(g(K_i))$$

where  $g$  represents the dependence from the neighbors. While  $g$  maybe linear or non-linear, the function  $f$ , representing the mode of updating, will typically be non-linear and/or probabilistic. In spite of this, the evolution of the Swarm only in special cases can be modeled by standard *non-linear dynamics* as studied via nonlinear differential equations (as we have seen in the previous section).

In regard to *optimization*, the Elementary Swarm can be easily designed to be an optimizer. In fact, if the updating is sequential, and  $f$  and  $v_i$  are chosen appropriately, one obtains a simple PSO system.

Thus, the Elementary Swarm describes a simple but powerful system capable, in principle, of self-organizing, and to produce complex structures and optimal solutions. On the other hand, even the Elementary Swarm is more general than these properties; a Swarm is not restricted by the notions of optimization or non-linear dynamics (and self-organization or complexity tied to non-linear dynamics). All these can be properties of the swarm but none is a requirement for SI. To find out what SI can do that is beyond what we have described so far, we must look at the computational capabilities of swarms. And for this we need to look at cellular-computing methods since standard-mathematics methods are ill suited to deal with computation.

### Cellular-Computing Methods

Some of the first studies in SI were based on computational models, and, more precisely, on distributed algorithms applied to robotic units [44,45,46].

In these computational models, the swarm was developed as an evolution from a distributed system of processors, as follows. In *distributed computing* the algorithms are designed for a ‘static’ set of processing units, where, ‘static’ is meant literally as ‘not moving’. For illustration, if a set of CPU’s, computing in a distributed way, via wireless communication, started moving around, this system would look very much like a robotic swarm. In fact, referring to the intuitive notion of SI (Definition 1), all points of the SI definition would be satisfied by such a dynamic, distributed computing system provided the CPU’s had, in some sense, limited capabilities.

The main point is that this distributed computing swarm differs from the robotic, and non-robotic, swarms described in the previous sections (Sects. “[Standard-Mathematics Methods](#)”–“[Limitations of Standard-Mathematics Methods](#)”) in that the *intelligent task of the swarm is now seen as a ‘computation’*. And this focus on computation leads us now to consider the other broad set of techniques used in SI research, i. e. techniques based not on standard-mathematics but on *cellular computing*.

Cellular computing differs qualitatively from the standard Von Neumann computing architecture. The latter is based on one complex processor that sequentially performs, at each time step, a single complex task. In contrast, in *cellular computing*, a very large number of simple processors (cells) are the units of computation. They compute (typically) in parallel with local connections between cells. The qualification ‘simple’ can be made precise by requiring, e. g., each cell to be a ‘finite state’ machine. Cellular automata are the most obvious examples of cellular computing systems but cellular computing applies to many other systems as well [47].

By definition, then, cellular computing contains several of the features of SI. The Elementary Swarm of Definition 2 can be regarded as performing a cellular computation. And in fact, cellular computing has been used extensively in A-Life studies, including systems with strong relation to SI [34]. Cellular computing systems offer SI something that the SI systems described in the ‘standard-mathematics’ Sects. “[Standard-Mathematics Methods](#)”–“[Limitations of Standard-Mathematics Methods](#)”) lack, i. e., a clear characterization of *intelligent task*.

### Intelligence as Universal Computation

Intelligence is an ambiguous concept, escaping a unique definition [48]. By identifying ‘intelligence’ with ‘computation’, the concept is restricted, but, at the same time, it can be made precise. In fact, in SI we define intelligence unambiguously as the ‘*ability to carry out universal computation*’.

Universal computation (or universality) is the property of a computer system (or language) which, with appropriate programming, can be made to perform exactly the same set of tasks as any other computer system (or language).

Universal computation (i. e., the ability to emulate a universal computer), is essentially the limit of any model of computation (Church–Turing thesis) [49]. It was first proven by Turing in 1936 that no system can ever carry out explicit computations more sophisticated than those carried out by a Turing machine. Subsequently, universal-

ity has been found to be a widespread property of many cellular computing systems [50].

One of the first cellular computing systems shown to be capable of universal computation, is Conway's game of life [51]. This CA is also the prototypical example of A-life systems. And it is also an example of the strong connection between universal cellular computing and bio-inspired systems.

More recently, a large number of simple cellular computing systems have been found to be capable of universal computation [50]. Many of these systems are CA, or related systems, using very simple rules of evolution with local interactions. And so they are useful starting points for modeling SI.

In particular, cellular computing is the most appropriate to endow the swarm with the property of *unpredictability*. The latter property was an original motivation for SI [1,2] and it is crucial in the task of escaping detection by a predator; it is also of importance in engineering swarms for strategic defense applications.

Unpredictability is almost a built-in property of cellular computing systems because, if one observes the rules of evolution in their raw form, it is usually almost impossible to tell much about the overall behavior they will produce.

### Relations to Standard-Mathematics Methods

SMm cannot provide the swarm with the element of universal computation, which we have taken as the working definition of 'intelligence'. The only way would be to make each unit a von Neumann (i. e. standard) computing system. In a sense, this violates the notion of Swarm, since in a Swarm, by definition each unit must be 'simple'. (This point will be further clarified in Sect. "Swarms of Intelligent Units"). On the contrary, the main advantage of cellular computing systems over standard mathematics systems is the possibility of universal computation by simple units.

For this reason CCm are the natural paradigm for the understanding and designing SI systems, in spite of the fact that the approach to SI based on cellular computing has so far produced fewer so-called SI applications than the approach based on SMm. Indeed, SMm have basic limitations for modeling SI. This is because the use of SMm tends to restrict the range of tasks performable by the Swarm. And this happens because SMm typically solve problems by specifying *constraints*, i. e., conditions to be satisfied by the solution, e. g., by specifying equations. But most computational problems cannot be solved in this way.

The optimization methods described in the previous sections (PSO, ACO, etc.), illustrate the point. In these iterative methods, the key issue is what kind of changes should be made at each iteration step. Starting from a random pattern, at each step a change is made to get the pattern closer to satisfying the constraint(s). Since direct methods (e. g., gradient descent) rarely work as the pattern gets stuck into local minima, randomness in updating is added. In this way, larger portions of the solution space are sampled. The larger the changes made, the faster one can potentially approach a global minimum, but the greater the chance of overshooting. The result is that no iteration technique of this type can guarantee a solution to general combinatorial optimization problems. As we have seen, the swarm optimization methods (ACO, PSO) rely on heuristics to adjust the search and obtain often (non-optimal but) satisfactory solutions. But, in general, for the great majority of combinatorial optimization problems (e. g., the Traveling Salesman Problem [52]), no polynomial upper bound on the time complexity has been found so far. And this happens in many problems whose solution is sought by using randomness to satisfy the imposed constraints. As an example, a set of identical balls cannot be shaken into an ordered, closed-packed configuration. With extremely high probability, they lock into some configuration or another, not the optimal (close-packing) one.

This fact has *important implications* for SI. What it says is that no matter how much randomness is added to the system, it may never evolve to reach the solution specified by the constraints. Although, ultimately, constraints can be set up as a way of specifying algorithms, and hence computing, it is far simpler to specify algorithms via rules of evolution, as it is done in cellular computing.

The conclusion is that methods based on constraints and other SMm are not ideally suited for systems evolving with great complexity, and in particular they are not suitable for universal computation. Thus, if SI is to be a framework for (biological or engineered) swarms to carry out 'intelligent tasks' with the greatest generality, a methodology that allows for the swarm to carry out universal computation is necessary. To this aim, CCm are the most suitable.

Unfortunately, although CCm have many advantages over SMm for modeling SI, they address only three of the four key elements of the notion (Definition 1) of SI ('intelligence', 'collectively', 'randomly', 'unknowingly'), leaving out the element of 'randomness'. Generally, CCm operate deterministically and do not include 'randomness', as, e. g., 'swarm optimization' systems do. But this does not have necessarily to be the case. The issue is addressed in the next section.

## Randomness in Swarm Intelligence

Randomness is a key element in the notion of SI (cfr. Definition 1 and Definition 2). Examples from biology justify this requirement. Randomness is not easily quantified precisely but, whatever the form and measure chosen, the point is that for swarms some form of randomness is necessary – otherwise they would fail to be models for analyzing a large class of biological systems. But what kind of randomness is essential to model these biological systems?

Randomness in the number and type of agents is not important – the agents could be strictly identical and remain in the same number. Randomness in the initial conditions is not essential either. Many swarms evolve from regular initial conditions into highly complex and random patterns. Randomness of external input from the environment is not always present, and it is certainly not a requirement for biological swarm behavior.

What about the randomness artificially added to the units, as in swarm optimization?

The randomness added to the units in PSO or ACO algorithms is modeled as originating from the random behavior of each unit. This is a plausible assumption in relation to biological systems. But the swarms in PSO and ACO are updated in an orderly (non-random) way, typically sequentially (there are also some parallel implementations [13]) whereas, in biological systems, the units update in a disordered, random fashion.

And it is this type of randomness that is both necessary in any biologically relevant model of swarms and sufficient to provide many (but not all) of the advantages of randomness in solving swarm engineering problems.

The conclusion is that the only randomness that is truly essential for SI is randomness in the times of operation of the units. Each unit has its own clock, not synchronized with other units' clocks. Other types of randomness in the behavior of the units or the environment may be required to solve specific problems, but randomness in times or operation is necessary for any biologically realistic model.

Interestingly though, many applications so far considered in the area of SI do not yet include this randomness in the models. We have already mentioned that typical optimizing swarms update sequentially; and CA systems operate largely in parallel, i. e. synchronously. Synchronous or sequential operations are by far the most common updating modes in either SMm or CCm.

## The Implicit Assumption of Asynchrony Irrelevance

As noted, it is a basic fact that biological agents, apart from exceptional cases, do not operate synchronously (or se-

quentially) in groups. It is also a fact that people in social groups do not operate synchronously, or sequentially. If SI is supposed to model biological and social swarms, SI must be based on models that do not operate synchronously or sequentially [53].

And if biological swarms are capable of solving problems (including optimization) without synchrony (nor sequentially), as they do, then models that imitate those swarms should operate asynchronously (not sequentially).

But, as noted, the main modeling paradigms for bio-inspired algorithms, standard-mathematics and cellular computing, are either essentially sequential or synchronous.

An example from SMm is the solution of partial differential equations: they operate synchronously on every point (clearly seen in solving them numerically and iteratively). This unrealistic use of differential equations in biological processes has been pointed out, e. g., in the problem of morphogenesis [54]. The Turing diffusion-reaction model [55], being based on differential equations, implies synchronicity and central control, hence it is physically not realistic for a scale of the order of 100 cells.

In fact, synchronicity leads to realistic models only whenever the spatio-temporal resolution is high, as, e. g., for phenomena typically studied in physics. But when the units studied are complex or few enough to have a less fine spatio-temporal resolution, as in biology or human societies, synchronicity is not realistic, as it is obvious by observation.

Thus, in using synchronous methods for biological or human societies, implicitly a strong assumption is being made, i. e., that the synchronously (or sequentially) and non-synchronously (and not sequentially) obtained solutions would coincide.

But this assumption has no validity. In fact, it has been shown, for example, that CA, when running in synchronous and non-synchronous ways, normally produce totally different results [56]. This has been noted already in the nineties [57] in A-Life studies. In [57] two well-known CA were compared: Conway's "game of life" [52], and the Immune Network model. The former is a 2-dimensional CA capable of universal computation when run synchronously. But the behavior is totally different when run without synchrony: the Game of Life stops producing complex patterns and converges to a fixed point.

The Immune Network model is asynchronous and the Game of Life synchronous. The crucial factor in the different behavior of the two systems was identified as the synchronous vs. asynchronous updating. In fact, it was concluded that, in this case, asynchrony induces stability



in CA. This agrees qualitatively with studies in standard-mathematics [58].

In conclusion, the assumption that asynchrony makes no difference has been found not to be valid. Hence, asynchronous systems must be studied as such, not by using synchronous models. Moreover, different types of asynchrony yield different results, as discussed in Sect. “Asynchronous Swarms”.

## Asynchronous Swarms

Several cellular computing studies in the nineties [37,59] led to a variety of results emphasizing the role that different types of asynchrony play in the results. Studying asynchronous systems is complicated because, among other things, deviation from synchronicity, i. e., from the mode of updating all units in parallel at each time step, may occur in several different ways. For example, sequential updating and random updating are both asynchronous but very different.

### Types of Asynchrony

Unfortunately, there is no standard vocabulary for the various types of asynchrony. Thus we use the following classification to describe the possible types of asynchrony.

Consider an updating cycle (UC), i. e., the time interval at the end of which all units have been updated at least once. Eight types of UC can be identified by the presence or absence of any of following three properties: (S) *Synchronicity*: more than one unit may update at each time

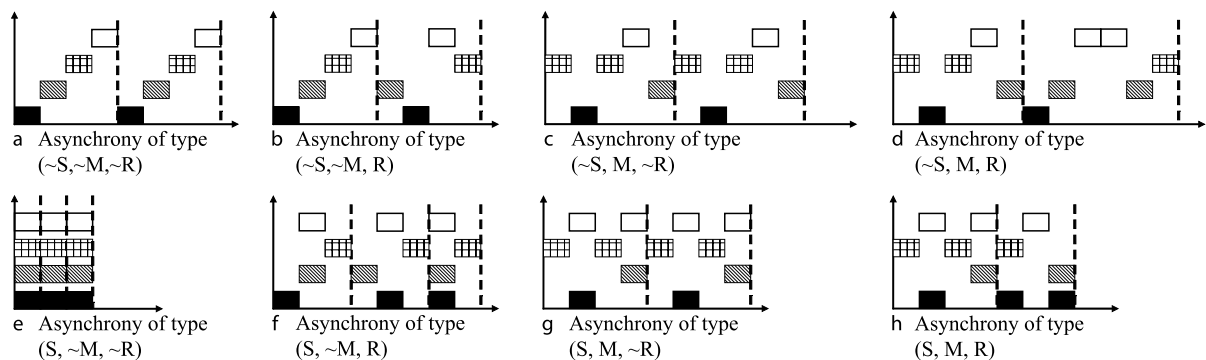
step; (M) *Multiplicity*: any unit may update more than once in each UC; (R) *Randomness*: the updating order varies randomly for every UC (see Fig. 5).

These eight basic types of asynchronous updating, can be further specialized. For example, if all three properties are absent ( $\sim S, \sim M, \sim R$ ), the updating is sequential. But the sequential updating order of the units can be fixed in different ways. Studies of CA have proven that the behavior differs markedly not only for the eight types of asynchrony, but even among different sequential ordering [37,56].

In [56] the  $(S, M, \sim R)$  form of updating has been applied to describe processes where each unit has independent clocks but the clocks have a fixed, non random frequency. This type of asynchrony is considered a good model for forest ecosystems, fire spread, and other natural and artificial systems. The results are very different when updating of the type  $(\sim S, M, R)$ ,  $(\sim S, \sim M, R)$ , or sequential  $(\sim S, \sim M, \sim R)$  are applied to the same system.

In conclusion, the crucial point is that [56] the exact manner of updating can have a profound effect on overall system behavior. The implication of this is that when comparing models of natural systems or artificial multi-agent systems it must be stated which updating scheme has been used, otherwise meaningful comparison between different studies may not be possible.

In particular, returning to swarm optimization, one may ask whether in tasks such as finding the shortest path, it is realistic to apply to natural systems (such as in-



Swarm Intelligence, Figure 5

a. Asynchrony of type  $(\sim S, \sim M, \sim R)$ ; b. Asynchrony of type  $(\sim S, \sim M, R)$ ; c. Asynchrony of type  $(\sim S, M, \sim R)$ ; d. Asynchrony of type  $(\sim S, M, R)$ ; e. Asynchrony of type  $(S, \sim M, \sim R)$ ; f. Asynchrony of type  $(S, \sim M, R)$ ; g. Asynchrony of type  $(S, M, \sim R)$ ; h. Asynchrony of type  $(S, M, R)$ ; Illustration of the 8 types of updating, according to Synchronicity, Multiplicity, and Randomness. Four units are represented by rectangles with different patterns, from black (bottom) to white (top). The horizontal axis measures time steps in units equal to the base of a rectangle. The vertical dashed line indicates the end of an updating cycle (i. e., all units have updated at least once). The label below the horizontal axis specifies the type of updating ( $\sim$  means ‘not’). The standard ‘parallel’ and ‘sequential’ updating are, respectively,  $(S, \sim M, \sim R)$ , Fig. 5e, and  $(\sim S, \sim M, \sim R)$ , Fig. 5a

sect societies) swarm models which are sequential (such as swarm optimization models) or synchronous (such as models based on differential equations). While these models work effectively as artificial swarms there is no proof that they apply to natural systems, which are asynchronous.

### Modeling Asynchrony by Synchronous Swarms

Because of the widespread use of synchronous methods in simulations of SI, one might wonder under what conditions a synchronous but stochastic model could be equivalent to an asynchronous one.

To answer this question, let us consider the two types of stochastic models most commonly used to model randomness in synchronously updated systems. The randomness may be included in (1) the possible outcomes of the updating function or (2) in the choice of the function applied to the updating.

Referring to the definition of Elementary Swarm (Definition 2) the two cases correspond to generalizing the updating function as follows:

$$\text{Case (1)} \quad \forall i \in N: v_i(t+1) = f(\mathbf{v}_{k \in K(i)}(t) \zeta)$$

where  $\zeta$  is a random variable.

$$\text{Case (2)} \quad \forall i \in N: v_i(t+1) = f_{(t)}(\mathbf{v}_{k \in K(i)}(t));$$

where  $P[f_{(t)} = f_\gamma]$  is the probability mass function of choosing  $f_{(t)} = f_\gamma$  out of a set of  $N_f$  possible functions  $\{f_\gamma; \gamma = 1, \dots, N_f\}$ .

Case (1) is typical of probabilistic CA, and it is also the method used in PSO. In these systems the state vector, at each time step, evolves according to a fixed rule which produce a new state vector from the previous one. The rule is based on the state of the neighbors of each unit and does not change from step to step but the outcome of the rule is probabilistic.

Case (2) is what is done for example in probabilistic Iterated Function Systems [23]. In probabilistic Iterated Function Systems a vector evolves via a set of maps (a map is a function whose domain and range coincide); at each time step a map is chosen, probabilistically, from a set of possible maps.

In either cases (1) or (2), the updating scheme fails to model the actual time evolution of natural systems not so much because the updating are applied synchronously but because the randomness is applied *collectively*, i.e. to all the units in the same way. On the other hand a synchronous algorithm realistically simulating independent random updating can be run as case (2) applied *individ-*

*ually* to each unit, as follows:

$$\text{Case (3)} \quad \forall i \in N: v_i(t+1) = f_{(t)i}(\mathbf{v}_{k \in K(i)}(t))$$

where  $P[f_{(t)1} = f_{\gamma 1}, f_{(t)2} = f_{\gamma 2}, \dots, f_{(t)N} = f_{\gamma N}]$  with  $f_{(t)i} \in \{f_\gamma; \gamma = 1, \dots, N_f\}$  is the joint probability mass function of each unit  $i$  updating, at time  $t$ , according to the function  $f_{(t)i}$ .

In the simplest embodiment of case (3), the set of possible updating functions consists only of the identity and of another function  $f$ , with probabilities  $p$  and  $(1-p)$  respectively (i.e., a Bernoulli process). In such a case, every unit, at each time step, either does not update, with probability  $p$ , or updates according to the function  $f$ , with probability  $(1-p)$ . Running this algorithm synchronously is equivalent to asynchronous independent updating of the units in a random way – a realistic description of a random swarm. So, under these independently stochastic conditions, running a simulation synchronously, represents correctly the physical asynchronous updating of the swarm units. On the other hand, this does not change the fact that different results are obtained when using this random updating (whether simulated with stochastic synchrony or not) instead of synchronous or sequential updating.

### Local Synchrony and Self-Synchronization

Another approach to dealing with the problem of random updating by the swarm units, is to explore the possibility of self-synchronization. If the swarm can self-synchronize, then all the results for synchronous swarms could be applied.

To look into this, let us return to the classification of the types of asynchrony, i.e., the SMR classification above. If the SMR properties are applied to blocks of units rather than individual units, the resulting updating orders are referred to as '*locally synchronous*'.

CA with cells organized into blocks have been investigated [60]. These CA relax the normal requirement of all cells having the same update rule. Cells within a block are updated synchronously, but blocks are updated asynchronously. They experimented with different SMR types of asynchrony and concluded that synchronous and asynchronous CA can be evolved with equivalent computational properties, but CA of the asynchronous type may require a larger number of cells [60]. Another study [61] has shown cases in which local synchronization can lead to the same outcome as with global synchronization. But how can local synchronization be achieved?

A number of schemes have appeared in which the order of updating depends on local interactions and leads

to local synchronization. In effect, what local interactions (or constraints) can do is to force a unit to wait to update until others are ready, and so this creates a local synchronization. An asynchronous CA model that can behave as a synchronous CA has been demonstrated [62]; it functions by the addition of extra constraints on the order of updating, effectively providing a type of local synchronization. Whether these methods of self-synchronization may in some cases result in realistic models of natural systems of SI remains an open question.

### The Natural Asynchrony of Swarms

We have seen that the implicit assumption of asynchrony-synchrony equivalence must be rejected and that different types of asynchrony give different results. But what type of asynchrony is most relevant to SI? There is no easy answer. For example, ants work and rest; active and resting periods have an aperiodic pattern for individual ants, but for the whole colony there are synchronized periodic patterns of active and resting periods.

In spite of the difficulty of finding a clear cut answer to the question of the natural mode of SI updating, from observations of biological systems, and from local synchronization models, it may be plausible to assume that the essential form of asynchrony in SI is the randomness in the working of the individual clocks, as argued in Sect. “Randomness in Swarm Intelligence”; hence the SI asynchrony must be characterized by the presence of all three asynchrony properties, i. e., (SMR).

In conclusion, at this stage of our discourse, the Swarm remains defined as in Definition 2, qualified by SMR asynchronous updating, which hereinafter we call ‘*natural*’ asynchrony. Note that stochastic synchronous simulations of this model can also be carried out as, e. g., in case (3) above.

### The Realization of Asynchronous Swarms

So far we have established the importance and type of asynchronous models in SI, but what SI investigations using asynchronous swarms are there?

As noted in Sect. “Asynchronous Swarms”, research in asynchronous models is still very limited, relative to synchronous models, and this in spite of the fact that the very first models of SI were all asynchronous, using SMm based on finite differences [46]. Explicit updating schemes in finite difference methods can also be regarded as parallel CA, thus belonging to both SMm and CCm. Investigations of asynchronicity in finite difference methods are not common [58]. Examples include a non-linear updating rule was based on a linear relation between two neighboring

units [45]. A gradient type of swarm updating, was also proposed in modeling morphogenesis [54].

For swarms updating with ‘*natural*’ asynchrony, i. e., according to (SMR), a recent study [58] gives a proof of convergence to the same state as by using synchronous or sequential iterations. It was also shown that, under certain conditions, the (SMR) asynchronous updating leads to convergence while synchronous updating does not. This is another example of the advantages of randomness in allowing the swarm to reach a fixed state.

At the end of Sect. “Cellular-Computing Methods” we concluded that CCm have, for SI modeling, many advantages over SMm. The most crucial advantage is the possibility of universal computation which we took as the definition of intelligence for SI. We also noted, however, that studies based on CCm which include randomness are scarce. We described a few in Sect. “Randomness in Swarm Intelligence”, especially in discussing the qualitative differences with synchronous CA and in relation to mechanisms of local and self-synchronization.

Generally these studies model relatively trivial phenomena but cannot model nontrivial phenomena such as universal computation. In fact, there are very few studies of universal computation in asynchronous CA. Significant advances have been made only recently. The first attempts were made by simulating a synchronous CA on an asynchronous CA [63] after which a synchronous model, as a Turing Machine, was simulated on the synchronous CA. However, this asynchronous CA is, in practical realization, synchronous.

Improved asynchronous CA do not rely on global synchronization but conduct asynchronous computation directly by simulating Delay-Insensitive circuits, i. e., circuits in which delays of signals do not affect the correctness of the circuit operation [64]. This method essentially uses local synchronization with undetermined exact timing between transitions. In this way an asynchronous CA, with a hexagonal cell structure, capable of universal computing has been realized [65]. Although relying on local synchronization, this type of asynchronous CA can mimic natural phenomena as, e. g., phenomena that rely on chemical reactions which occur only when the right molecules are available in the right positions at the right times.

More recently a computation-universal and construction-universal asynchronous CA has been designed [66] and used to implement self-reproducing [67,68] machines. Besides computational universality, construction universality is important in SI because it allows the swarm to be hardware reconfigurable, an important characteristic of many biological systems.

We note that the recent interest in asynchronous CA stems not directly from SI but from nanotechnology. In fact nanocomputer architectures with asynchronous updating may reduce heat dissipation, an important limiting factor in scaling down the size of computing chips [64]. In this respect, it is likely that SI concepts will play a major role in nanoscale systems.

In conclusion, the recent realization [66] of universal asynchronous CA is a major step toward the realization of true SI.

### Characteristics of Swarm Intelligence

The demonstration of universal computation in asynchronous CA, amounts to a validation of the concept of SI. In fact we can now combine universal computation with the Elementary Swarm definition (Definition 2) to quantify the intuitive definition of SI (Definition 1), as

**Definition 3** SI is the study of *universal cellular-computing systems updating with natural asynchrony*.

Here ‘natural’ means SMR-asynchrony (see Sect. “Asynchronous Swarms”) or updating randomly in parallel, as in case (3) of Sect. “Asynchronous Swarms”. We call it ‘natural’ since it models the natural mode of updating of typical biological swarms and human societies.

A few remarks about Definition 3. The four elements of the intuitive notions (Definition 1) are made precise by Definition 3: ‘collectively’ and ‘unknowingly’ are inherent in the structure of cellular computing; ‘intelligence’ is in universal computation; and ‘randomness’ is in the natural asynchronous operation.

Definition 3 deals with CCm and may appear to exclude SMm in SI; but this is not the case. In fact, many SMm, as used in SI, can be regarded as special cases of cellular computing methods, as, e. g., are swarm optimization and iterative methods in finite differences.

Also Definition 3 does imply that every SI system must be capable of universal computation; what Definition 3 does is to establish a focus of attention for the SI area of studies and at the same time give a precise and realistic meaning to the kind of ‘intelligence’ aimed at in SI, rather than the often vague and exaggerated meanings given in the popular literature.

Definition 3 also indicates how SI becomes of relevance beyond biological systems and robotics. In fact SI will likely be an important concept in the future of computation. At very small scales, time delays between computational components cannot guarantee synchrony; the various components must have independent clocks, thus beginning to resemble the operation of a biological swarm.

So, swarms are likely to be studied extensively in connections with nanocomputing.

The above arguments bring up the question as to whether SI is nothing more than asynchronous cellular computing. The answer is that designs of asynchronous CA, as investigated in computer engineering, are generally not models of SI. A first, basic reason is that, with one recent exception [66], asynchronous CA, do not update ‘naturally’. A second, more fundamental reason, is that in most SI studies (both in natural and technological systems) **the units are dynamic**. When the units of an asynchronous CA are made mobile, a different, and more complex set of problems need to be solved due to the changing neighborhoods of each cell. These issues of dynamic reconfigurations of cells have not been addressed in asynchronous cellular computing designs and are likely to remain outside the scope of research aimed at improved computer architectures. The computational problems arising from dynamically reconfiguring cells are central in SI. We address this issue next.

### Dynamics in Swarm Intelligence

In the definitions of SI and Swarm given so far (Def. 1., Def. 2., Def. 3.), there has been no mention of the dynamics of the units. But a general characteristic of the units of a swarm is that almost invariably they are mobile. In fact we have already discussed the dynamic nature of swarms in Sects. “Biological Systems” and “Robotic Systems” in relation to biological and robotic swarms. The reason we have so far omitted this dynamic character of the units from the progressively more precise definitions of SI is simply for clarity of exposition: if the dynamics is introduced after all the other elements of SI have been defined and quantified it is easier to single out its real importance.

From Definition 3, with appropriate specializations, all aspects of SI considered so far can be included in a common core of studies. The most general notion of SI is in fact that of *universal computation carried out with ‘natural asynchrony’ by a cellular computing system*. But we should add, *whose cells are, in general, mobile units*.

The latter qualification would be unnecessary if the description of the dynamic state could be included among the state variables of the cell. But this is not possible in computing cells since the computation depends on neighbors that change their locations. The fact that *dynamic cellular computing systems* (also called *cellular robotic systems*) are not equivalent to (and very hard to simulate by) cellular computers has been emphasized since the very beginning of SI studies [44]. Thus, we conclude by stating a definition of SI which, while remaining grounded in the

intuitive ideas (Definition 1), includes all the concepts discussed quantitatively in this article.

**Definition 4** SI is the capability of universal computation carried out with ‘natural’ asynchrony by a dynamic cellular computing system.

As we have seen, studies in the area of SI so far have been concerned with models of collective behavior which, to some limited degree, approach SI as defined above. Even though, to date, no system with SI (Def. 4.) has been built (or designed), significant progress has been made and, from what we have seen, it is reasonable to expect that it can be done.

In fact, although there is not yet proof of universal computation carried out with ‘natural’ asynchrony by a *cellular robotic system* (i. e., a dynamic cellular computing system), the recent proof [66] for ‘static’ cellular computers indicates that this may be possible in the near future. Other future perspectives for SI are discussed in Sect. “Future Directions”.

### Unpredictability in Swarm Intelligence

We are now in a position to consider an aspect of SI which has been inherent to the concept of SI from its inception, i. e., the unpredictability [1,2] of the Swarm.

The unpredictability of the Swarm agrees with the common intuition that it is usually difficult to predict what a program will do by reading its code, and the more so the lower the level of the language used by the program. More precisely, a Swarm, like other universal computers, may be impossible to predict in the sense that even if one knows the rules of evolution and an initial state, it can still take an irreducible [50,69] amount of computation to actually predict future states. Furthermore, the unpredictability of the Swarm is of a more general character than that of any universal computer because of the randomness inherent in its evolution and because of its dynamics.

The unpredictability of a Swarm by a Von Neumann universal computer has been argued in [1,10] on the basis of its dynamics. The unpredictability of a Swarm by a cellular automaton has also been discussed [1,10]. Although unpredictability is difficult to quantify, it is generally engineered by adding randomness to the system as in camouflage and cryptography. Also, in animals, randomness and dynamics are used by a herd to avoid predators by becoming unpredictable. And so in team sports, such as soccer, unpredictability by the opponent is usually achieved by a combination of randomness and dynamics.

Therefore, intuitively we may conjecture that among systems capable of universal computation, a Swarm, be-

cause it operates with randomness and dynamically, would be *the least predictable*.

### Swarms of Intelligent Units

We can now consider collections of intelligent units, i. e., such that each unit is capable of universal computation. These seem excluded from the definitions of SI given so far – a key characteristic of SI is that intelligence is an emergent property, happening only above a certain critical number of units, and not a property of any of the individual units. On the other hand, some of these groups are often included in broad considerations of SI [4] as applied to human societies. Under what conditions can these groups be regarded as swarms?

A simple answer to this question runs as follows. Consider the special case of Definition 4, when the cells are not finite state machines but universal computing units. As long as the task at hand cannot be accomplished by a single unit, but only by more than a critical number of units, the system operates as a swarm. That this can be the case is supported intuitively and from computational considerations, as follows.

Intuitively, we may refer back to the example of the free market model mentioned as illustration of SI at the beginning of the article. Each individual contributes only as a trader; the ‘computation’ of the market price is done by the swarm collectively and could not be done by any individual agent. The point is that each unit, albeit ‘intelligent’, uses only a fraction of its capability, i. e., the trading ability, thus operating, effectively, in a restricted, ‘non-intelligent’ capacity.

Computationally, we have noted in discussing unpredictability that, in spite of theoretical computational equivalence among universal computers, the capability for one universal computer to predict another is limited. And, in fact, at the end of Sect. “Swarms of Intelligent Units”, we put forth the conjecture that a Swarm is the least predictable universal computer.

But even if this conjecture were not true, it still makes sense to think of a case when a Swarm of universal computers is unpredictable by any of the units comprising it, as it has been argued for the case of units capable of universal computation with Von Neumann architecture [1,10]. In this sense we may think of a swarm of universal computers as being more capable than any one of its units, in spite of the fact that the Swarm and any of its units are computationally equivalent. For these reasons, it makes sense to apply, under appropriate conditions, the notion of SI also to human societies.



We may give then a more complete definition of Swarm by adding the characteristic of unpredictability of the Swarm by its units.

**Definition 5** SI is the capability of universal computation carried out with ‘natural’ asynchrony by a dynamic cellular computing system, none of whose cells can predict the computation done by the Swarm.

The latter specification is obviously redundant for common cellular computing systems but it is useful to exclude from SI trivial cases of human social activities; for example, activities of human groups whose association cannot be proven to solve a problem that could not have been solved by any individual alone.

Many popular interpretations of SI have appropriated the label SI to refer to almost any trivial human group activity, such as brainstorming. In such activities often there is no way of proving that the final output of the group could not have been predicted by one of the members of the group. Ultimately, however, these considerations require an understanding of the relation between human thinking and computation, and thus fall beyond the scope of this article.

Similarly, beyond the scope of this article and SI, fall studies of multiagent systems in Artificial Intelligence [70]. In the Artificial Intelligence area, the emphasis on multi-agent systems is in finding decision algorithms, i. e. ‘agents’, for open environments in which these agents must operate robustly and rapidly, i. e. ‘intelligently’. Generally, however, the problem of collective decision making, organization theory, distributed reasoning and distributed Artificial Intelligence is typically beyond the scope of SI.

SI deals with human groups only when they operate at a low level of intelligence. Definition 5 includes human groups of individuals that operate under restrictions which the Swarm can overcome. An example is that of groups of individuals each with a limited computation device connected wirelessly to neighbors as in a cellular computing system. With appropriate algorithms such a Swarm could compute universally while the individual device cannot. This concept is applicable in, e. g., defense operations or emergency mass evacuations strategies.

In conclusion, Definition 5 embodies the intuitive definition of SI (Definition 1) and indicates why SI methods can be used to solve problems not solvable by traditional methods. Besides having all the properties of universal CA, the Swarm operates, as natural systems do, by independent clocks with no centralization and can be designed to be dynamic and unpredictable by any system including any its own units.

## Future Directions

The field of SI is still less than 20 years old and it is in a formative phase, with SI researchers engaged in a broad range of disciplines. During these formative years the ‘conversation’ about SI has followed several strands around some common themes with various emphases ranging from speculative inquiries to practical interests, as we have seen in this article.

The meaning of the term SI has tended to broaden, covering now many areas, apparently only weakly related by the some intuitive notions. We have seen, however, how all these SI ideas have a ‘center of attraction’ in a basic concept of SI that can be made precise and quantified (Definition 4 or 5), and thus used to provide unity, continuity, and boundaries, thus preventing the area from broadening to the point of being unable to sustain an effective research community.

With this perspective, SI can also be seen as having an ultimate theoretical goal for the practical realization of engineered Swarms, whether robotic, biological or simply computational.

Practically, the goal of SI will remain two-fold: to provide models to explain biological societies and to engineer algorithms and devices with capability beyond those of traditional technologies. It will continue to include Swarm Robotics and bio-inspired algorithms such as swarm optimization methods.

But although commonly regarded as a typical example of bio-inspired technology, SI applications are likely to go beyond bio-inspired systems. We can see this if we consider how nature-inspired technologies have evolved.

Science discovers laws of nature, and technology makes inventions using those laws, often together with design ideas also derived from nature. Thus, for example, laws of physics and designs inspired by crystal structures are now applied to nanotechnologies; similarly, laws of biology and designs inspired by genetic configurations are applied to make artificial organisms in biotechnology. More recently a new kind of science [50] is discovering the laws and designs of computing machines as though they were natural systems, and these discoveries are likely to be used to invent new software algorithms and hardware implementations of those algorithms. SI, besides being bio-inspired, can be said to be inspired also by this new science, which, in some respects, can be more general than biology.

In this perspective, SI will evolve into the study of what amounts to be very powerful computing systems. Designing the simplest of these, i. e., the simplest universal dynamic cellular-computing system updating with natural

asynchrony, is an example of a future theoretical and practical challenge for SI. More immediate, future applications can be extrapolated from the examples given throughout the article.

## Bibliography

### Primary Literature

- Beni G, Wang J (1989) Swarm Intelligence in Cellular Robotic Systems, *Proceed. NATO Advanced Workshop on Robots and Biological Systems*, Tuscany, Italy, June 26–30
- Beni G, Wang J (1989) Swarm Intelligence. *Proc. 7th Ann Meeting of the Robotics Society of Japan*, pp 425–428 (in Japanese)
- Bonabeau E, Dorigo M, Theraulaz G (1999) *Swarm Intelligence: From Natural to Artificial Systems*. Oxford Univ Press, New York
- Kennedy J, Eberhart RC, Shi Y (2001) *Swarm Intelligence*. Morgan Kaufman, San Mateo
- Abraham A, Grosan C, Ramos V (2006) *Swarm Intelligence in Data Mining (Studies in Computational Intelligence)*. Springer, vol 34
- Engelbrecht AP (2006) *Fundamentals of Computational Swarm Intelligence*. Wiley, New York
- Dorigo M, Gambardella LM, Birattari M, Martinoli A (eds) (2006) *Ant Colony Optimization and Swarm Intelligence: 5th International Workshop, ANTS 2006, Brussels, Belgium, September 4–7, (2006), Proceedings (Lecture Notes in Computer Science)*. Springer, Berlin
- IEEE Swarm Intelligence Symposium, Honolulu, Hawaii, 1–5 April 2007. <http://www.computelligence.org/sis/2007/?q=node/2>. Accessed 6 Mar 2008
- Dorigo M (ed) *Swarm Intelligence*, ISSN: 1935-3812. Springer US (to begin publication Summer 07)
- Beni G (2004) From Swarm Intelligence to Swarm Robotics. In: Sahin E, Spears WM (eds.) *Swarm Robotics March, Revised Selected Papers, SAB 2004 International Workshop, Santa Monica, CA, 17 July 2004. Lecture Notes in Computer Science vol 3342*. Springer, Berlin, pp 1–9
- Goss S, Aron S, Deneubourg JL, Pasteel JM (1989) Self-organized shortcuts in the Argentine ant. *Naturwissenschaften* 76:579–581
- Dorigo M (1992) *Optimization, Learning and Natural Algorithms*. Ph D Thesis, Dipartimento di Elettronica, Politecnico di Milano, Milan (in Italian)
- Dorigo M, Stutzle T (2004) *Ant Colony Optimization*. MIT Press, Cambridge
- Olariu S, Zomaya AY (2005) *Handbook of Bioinspired Algorithms and Applications*. Chapman Hall/Crc Computer Information Science, Boca Raton
- Passino K (2004) *Biomimicry for Optimization, Control, and Automation*. Springer, London
- Reynolds C (1987) Flocks, Herds, and Schools: a distributed behavioral model. *Computer Graphics* 21(4):25–34
- Moore AV (2004) Information Flocking: Time-Varying Data Visualization using Boid Behaviors. *Proc Eighth Int Conf Inform Visual*, pp 409–414
- Eftimie R, de Vries G, Lewis MA (2007) Complex spatial group patterns result from different animal communication mechanisms. *Proc Nat Acad Sci* 104(17)
- Parker LE, Schneider FE, Schultz AC (2005) Multi-Robot Systems. *From Swarms to Intelligent Automata*, vol 3: *Proceedings from the 2005 International Workshop on Multi-Robot Systems*. Springer, Dordrecht
- Sahin E, Spears WM (2005) *Swarm Robotics. SAB 2004 International Workshop, Santa Monica, CA, July 17, 2004, Revised Selected Papers (Lecture Notes in Computer Science)*. Springer, Berlin
- Winfield A, Redi J <http://www.robocomm.org/2007/index.shtml>. Accessed 6 Mar 2008
- Sahin EL, Spears WM, Winfield AFT (Eds.) (2007) *Swarm Robotics. Second SAB 2006 International Workshop, Rome, Italy, September 30–October 1, 2006 Revised Selected Papers. Series: Lecture Notes in Computer Science vol 4433*. Springer, Berlin
- Baldassarre G, Trianni V, Bonani M, Mondada F, Dorigo M, Nolfi S (2007) Self-organized coordinated motion in groups of physically connected robots. *IEEE Trans Syst Man Cybern Part B: Cybernetics* 37(1):224–239
- Gazi V, Passino KM (2004) Stability Analysis of Social Foraging Swarms. *IEEE Trans Syst Man Cybern B* 34:539–557
- Balch T, Dellaert F, Feldman A, Guillory A, Isbell CL Jr, Khan Z, Pratt SC, Stein AN, Wilde H (2006) How AI and Multi-Robot Systems research will accelerate our understanding of social animal behavior. *Proc IEEE* 94(7):1445–1463
- Brooks R (1986) A robust layered control system for a mobile robot. *IEEE J Robot Autom* RA-2(1):14
- Johnson N, Galata A, Hogg DB (1998) The acquisition and use of interaction behavior models. In: *Proc IEEE Computer Society Conf Computer Vision and Pattern Recognition*, pp 866–871, IEEE Computer Society Press, Santa Barbara
- Curtis SA, Mica J, Nuth J, Marr G, Rilee ML, Bhat M (2000) *Autonomous Nano-Technology Swarm. Proceedings of the 51st International Aeronautical Congress, IAF-00-Q5.08*
- ANTS website: <http://ants.gsfc.nasa.gov>, Goddard Ants Team. Accessed 6 Apr 2008
- Brooks RA (1991) Intelligence without representation. *Artif Intell* 47:139–159
- Dorigo M, Tuci E, Groß R, Trianni V, Labella TH, Nouyan S, Ampatzis C, Deneubourg J-L, Baldassarre G, Nolfi S, Mondada F, Floreano D, Gambardella LM (2004) The SWARM-BOTS project. In: Sahin E, Spears WM (eds) *Proc. of the 1st Int. Workshop on Swarm Robotics. LNCS vol 3342*. Springer, Berlin, pp 26–40
- Mondada F, Pettinaro GC, Guignard A, Kwee IV, Floreano D, Deneubourg J-L, Nolfi S, Gambardella LM, Dorigo M (2004) SWARM-BOT: A new distributed robotic concept. *Auton Robots* 17(2–3):193–221
- Mondada F, Gambardella LM, Floreano D, Nolfi S, Deneubourg J-L, Dorigo M (2005) The cooperation of swarm-bots: Physical interactions in collective robotics. *IEEE Robot Autom Mag* 12(2):21–28
- Rocha LM, Yaeger LS, Bedau MA, Floreano D, Goldstone RL, Vespignani A (2006) *Artificial Life X Proceedings of the Tenth International Conference on the Simulation and Synthesis of Living Systems*. MIT Press, Cambridge
- Traub J (ed) *Journal of Complexity*. Elsevier [http://www.elsevier.com/wps/find/journaldescription.cws\\_home/622865/description#description](http://www.elsevier.com/wps/find/journaldescription.cws_home/622865/description#description)
- Nicolis G, Prigogine I (1977) *Self-Organization in Non-Equilibrium Systems*. Wiley, New York

37. Sipper M (1999) The Emergence of Cellular Computing. *IEEE Comput* 32(7):18–26
38. Kennedy J, Eberhart RC (1995) Particle swarm optimization. In: *Proc. IEEE int'l conf. on neural networks vol IV*, pp 1942–1948. IEEE service center, Piscataway
39. Kirkpatrick S, Gelatt CD, Vecchi MP (1983) Optimization by Simulated Annealing. *Science* 220(4598):671–680
40. Snyman JA (2005) *Practical Mathematical Optimization: An Introduction to Basic Optimization Theory and Classical and New Gradient-Based Algorithms*. Springer Science-Business Media, New York
41. Mogilner A, Edelstein-Keshet L, Bent L, Spiros A (2003) Mutual interactions, potentials, and individual distance in a social aggregation. *J Math Biol* 47:353–389
42. Murray JD (2002) *Mathematical Biology I: An Introduction*, 3rd edn., *Interdiscip Appl Math* 17. Springer, New York
43. Topaz CM, Bertozzi A (2004) Swarming Patterns in two-dimensional kinematic model for biological groups. *SIAM J Appl Math* 65(1):152–174
44. Beni G (1988) The concept of cellular robot. *Proc 3rd IEEE Symposium on Intelligent Control*, pp 57–61. Arlington
45. Beni G, Hackwood S (1992) Stationary waves in cyclic swarms. In: *Proc IEEE International Symposium on Intelligent Control*. August 10–13, 1992. Institute of Electrical and Electronics Engineers (IEEE), Glasgow
46. Beni G (1992) Distributed Robotic Systems and Swarm Intelligence. *J Robot Soc Japan* (in Japanese) 10:31–37
47. Sipper M (1997) *Evolution of Parallel Cellular Machines: The Cellular Programming Approach*. Lecture Notes in Computer Science. Springer
48. Gottfredson LS (1997) Mainstream Science on Intelligence: An Editorial with 52 Signatories, History, and Bibliography. *Intelligence* 24(1):13–23
49. Cooper SB (2003) *Computability Theory*. Chapman Hall/CRC, Boca Raton
50. Wolfram S (2002) *A New Kind of Science*. Wolfram Media, Champaign
51. Gardner M (1970) The fantastic combinations of John Conway's new solitaire game 'life'. *Sci Am* 223:120–123
52. Johnson DS, McGeoch LA (2003) The traveling salesman problem: A case study in local optimization. In: Aarts EHL, Lenstra JK (eds) *Local Search in Combinatorial Optimization*, pp 215–310. Wiley, Chichester
53. Huberman BA, Glance NS (1993) Evolutionary Games and computer simulations. *Proc Nati Acad Sci USA* 90:7716–7718
54. Liang P, Beni G (1995) Robotic Morphogenesis. *Proc Int Conf Robot Automat* 2:2175–2180
55. Turing AM (1952) The Chemical Basis for Morphogenesis. *Phil Trans Roy Soc London B* 237:37–72
56. Cornforth D, Green D, Newth D (2005) Ordered Asynchronous Processes in Multi-Agent Systems. *Physica D* 204(1–2):70–82
57. Bersini H, Detour V (1994) Asynchrony induces stability in CA based models. In: Brooks RA, Maes P (eds) *Artificial Life IV*. MIT Press, Cambridge, pp 382–387
58. Beni G (2004) Order by Disordered Action in Swarms. In: Sahin E, Spears WM (eds.) (2005) *Swarm Robotics, SAB 2004 International Workshop*. LNCS, vol 3342. Springer, Berlin, pp 153–171
59. Schonfisch B, deRoos A (1999) Synchronous and Asynchronous Updating in Cellular Automata. *Biosyst* 51:123–143
60. Sipper M, Tomassini M, Capcarrere MS (1997) Evolving asynchronous and scalable non-uniform cellular automata. In: *Proceedings of International Conference on Artificial Neural Networks and Genetic Algorithms (ICANNGA97)*. Springer, New York
61. Clapham N (2002) Emergent synchrony: simple asynchronous update rules can produce synchronous behavior. In: Sarker R, McKay RI, Gen M, Namatame A (eds) *Proceedings of the Sixth Australia–Japan Joint Workshop on Intelligent and Evolutionary Systems*. Australian National University, Canberra, pp 41–46
62. Nehaniv CL (2002) Evolution in asynchronous cellular automata. In: Standish RK, Abbass HA, Bedau MA (eds) *Proceedings of the Eighth Conference on Artificial Life*. MIT Press, Cambridge, pp 65–74
63. Nakamura K (1974) Asynchronous cellular automata and their computational ability. *Syst Comput Controls* 5(5):58–66
64. Lee J, Peper F, Adachi S, Morita K (2004) Universal delay-insensitive circuits with bi-directional and buffering lines. *IEEE Trans Comput* 53(8):1034–1046
65. Adachi S, Peper F, Lee J (2004) Universality of hexagonal asynchronous totalistic cellular automata. In: Chopard B, Hoekstra AG, Sloot PMA (eds) (2004) *Cellular Automata, 6th International Conference on Cellular Automata for Research and Industry*. LNCS, vol 3305. Springer, Berlin, pp 91–100
66. Takada Y, Isokawa T, Peper F, Matsui N (2006) Construction Universality in purely asynchronous cellular automata. *J Comput Syst Sci* 72:1368–1385
67. von Neumann J (1966) *Theory of Self-Reproducing Automata*. Univ Illinois Press, Champaign (edited and completed by Burks AW)
68. Langton CG (1984) Self-reproduction in cellular automata. *Phys D* 10:135–144
69. Wolfram S (1985) Undecidability and Intractability in Theoretical Physics. *Phys Rev Lett* 54:735–738
70. Weiss G (2000) *Multiagent Systems: A Modern Approach to Distributed Artificial Intelligence*. MIT Press, Cambridge

## Books and Reviews

- References [3, 4, 6, 10, 13, 14, 15, 48, 51] are the main books and reviews. The journal *Swarm Intelligence* [9] is the main source for new research results. The following provide additional general material related to Swarm Intelligence.
- Sahin E, Spears WM (2005) *Swarm Robotics: SAB 2004 International Workshop*, Santa Monica, CA, July 17, 2004, Revised Selected Papers (Lecture Notes in Computer Science). Springer, Berlin
- Dorigo M, Sahin E (2004) *Swarm Robotics—special issue editorial*. *Autonom Robot* 17(2–3):111–113
- Sipper M (2002) *Machine Nature: The Coming Age of Bio-Inspired Computing*. McGraw-Hill, New York
- Camazine S, Deneubourg J-L, Franks NR, Sneyd J, Theraulaz G, Bonabeau E (2001) *Self-Organization in Biological Systems*. Princeton Univ Press, Princeton

## Symbolic Dynamics

BRIAN MARCUS  
Department of Mathematics,  
University of British Columbia, Vancouver, Canada

## Article Outline

Glossary  
 Definition of the Subject  
 Introduction  
 Origins of Symbolic Dynamics: Modeling of Dynamical Systems  
 Shift Spaces and Sliding Block Codes  
 Shifts of Finite Type and Sofic Shifts  
 Entropy and Periodic Points  
 The Conjugacy Problem  
 Other Coding Problems  
 Coding for Data Recording Channels  
 Connections with Information Theory and Ergodic Theory  
 Higher Dimensional Shift Spaces  
 Future Directions  
 Bibliography

## Glossary

In this glossary, we give only brief descriptions of key terms. We refer to specific sections in the text for more precise definitions.

**Almost conjugacy** (Sect. “Other Coding Problems”) A common extension of two shift spaces given by factor codes that are one-to-one almost everywhere.

**Automorphism** (Sect. “The Conjugacy Problem”) An invertible sliding block code from a shift space to itself; equivalently, a shift-commuting homeomorphism from a shift space to itself; equivalently, a topological conjugacy from a shift space to itself.

**Dimension group** (Sect. “The Conjugacy Problem”) A particular group associated to a shift of finite type. This group, together with a distinguished sub-semigroup and an automorphism, captures many invariants of topological conjugacy for shifts of finite type.

**Embedding** (Sect. “Shift Spaces and Sliding Block Codes”) A one-to-one sliding block code from one shift space to another; equivalently, a one-to-one continuous shift-commuting mapping from one shift space to another.

**Factor map** (Sect. “Shift Spaces and Sliding Block Codes”) An onto sliding block code from one shift space to another; equivalently, an onto continuous shift-commuting mapping from one shift space to another. Sometimes called *Factor Code*.

**Finite equivalence** (Sect. “Other Coding Problems”) A common extension of two shift spaces given by finite-to-one factor codes.

**Full shift** (Sect. “Shift Spaces and Sliding Block Codes”) The set of all bi-infinite sequences over an alphabet (together with the shift mapping). Typically, the alphabet is finite.

**Higher dimensional shift space** (Sect. “Higher Dimensional Shift Spaces”) A set of bi-infinite arrays of a given dimension, determined by a collection of finite forbidden arrays. Typically, the alphabet is finite.

**Markov partition** (Sect. “Origins of Symbolic Dynamics: Modeling of Dynamical Systems”) A finite cover of the underlying phase space of a dynamical system, which allows the system to be modeled by a shift of finite type. The elements of the cover are closed sets, which are allowed to intersect only on their boundaries.

**Measure of maximal entropy** (Sect. “Connections with Information Theory and Ergodic Theory”) A shift-invariant measure of maximal measure-theoretic entropy on a shift space. Its measure-theoretic entropy coincides with the topological entropy of the shift space.

**Road problem** (Sect. “Other Coding Problems”) A recently-solved classical problem in symbolic dynamics, graph theory and automata theory.

**Run-length limited shift** (Sect. “Coding for Data Recording Channels”) The set of all bi-infinite binary sequences whose runs of zeros, between two successive ones, are bounded below and above by specific numbers.

**Shift equivalence** (Sect. “The Conjugacy Problem”) An equivalence relation on defining matrices for shifts of finite type. This relation characterizes the corresponding shifts of finite type, up to an eventual notion of topological conjugacy.

**Shift space** (Sect. “Shift Spaces and Sliding Block Codes”) A set of bi-infinite sequences determined by a collection of finite forbidden words; equivalently, a closed shift-invariant subset of a full shift.

**Shift of finite type** (Sect. “Shifts of Finite Type and Sofic Shifts”) A set of bi-infinite sequences determined by a *finite* collection of finite forbidden words.

**Sliding block code** (Sect. “Shift Spaces and Sliding Block Codes”) A mapping from one shift space to another determined by a finite sliding block window; equivalently, a continuous shift-commuting mapping from one shift space to another.

**Sofic shift** (Sect. “Shifts of Finite Type and Sofic Shifts”) A shift space which is a factor of a shift of finite type; equivalently, a set of bi-infinite sequences determined by a finite directed labeled graph.

**State splitting** (Sect. “The Conjugacy Problem”) A splitting of states in a finite directed graph that creates



a new graph, whose vertices are the split states. The operation that creates the new graph from the original graph is a basic building block for all topological conjugacies between shifts of finite type.

**Strong shift equivalence** (Sect. “[The Conjugacy Problem](#)”) An equivalence relation on defining matrices for shifts of finite type. In principle, this relation characterizes the corresponding shifts of finite type, up to topological conjugacy.

**Topological conjugacy** (Sect. “[Shift Spaces and Sliding Block Codes](#)”) A bijective sliding block code from one shift space to another; equivalently, a shift-commuting homeomorphism from one shift space to another. Sometimes called *conjugacy*.

**Topological entropy** (Sect. “[Entropy and Periodic Points](#)”) The asymptotic growth rate of the number of finite sequences of given length in a shift space (as the length goes to infinity).

**Zeta function** (Sect. “[Entropy and Periodic Points](#)”) An expression for the number of periodic points of each given period in a shift space.

## Definition of the Subject

Symbolic dynamics is the study of shift spaces, which consist of infinite or bi-infinite sequences defined by a shift-invariant constraint on the finite-length sub-words. Mappings between two such spaces can be regarded as codes or encodings. Shift spaces are classified, up to various kinds of invertible encodings, by combinatorial, algebraic, topological and measure-theoretic invariants.

The subject is intimately related to many other areas of research, including dynamical systems, ergodic theory, automata theory and information theory. Shift spaces and their associated shift mappings are used to model a rich and important class of smooth dynamical systems and ergodic measure-preserving transformations. These models have provided a valuable tool for classifying and understanding fundamental properties of dynamical systems. In addition, techniques from symbolic dynamics have had profound applications for data recording applications, such as algorithms and analysis of invertible encodings, and problems in matrix theory, such as characterization of the set of eigenvalues of a nonnegative matrix.

## Introduction

This article is intended to give a picture of major topics in symbolic dynamics. Section “[Origins of Symbolic Dynamics: Modeling of Dynamical Systems](#)” reviews the roots of symbolic dynamics in modeling of dynamical systems. Section “[Shift Spaces and Sliding Block Codes](#)”

lays the foundation by defining the kinds of spaces and mappings considered in the subject. Section “[Shifts of Finite Type and Sofic Shifts](#)” focuses on distinguished special classes of spaces, known as shifts of finite type and sofic shifts. Section “[Entropy and Periodic Points](#)” introduces the most fundamental invariants, periodic points and topological entropy. Sections “[The Conjugacy Problem](#)” and “[Other Coding Problems](#)” survey progress on the conjugacy problem and other classification/coding problems for shifts of finite type and sofic shifts. In Sect. “[Coding for Data Recording Channels](#)”, we present applications to coding for data recording. Section “[Connections with Information Theory and Ergodic Theory](#)” provides a link with information theory and ergodic theory. Finally, Sect. “[Higher Dimensional Shift Spaces](#)” treats higher dimensional symbolic dynamics.

While this article covers many of the most important topics in the subject, others have been omitted or treated lightly, due to space limitations. These include one-sided shift spaces, countable state symbolic systems, orbit equivalence, flow equivalence, the automorphism group, cellular automata, and substitution systems. References to work in these sub-areas can be found in the sources mentioned below.

For introductory reading on symbolic dynamics and its applications, beyond this article, one can consult the textbooks Kitchens [65] and Lind and Marcus [78]. There are also excellent introductory survey articles, such as Boyle [23], Lind and Schmidt [79], and S. Williams [132]. In addition there are very good expositions which focus on other aspects of the subject. These include Beal [11], which focuses on connections between symbolic dynamics and automata theory, the lecture notes Marcus–Roth–Siegel [83], which focuses on constrained coding applications, and Immink [54], which focuses on applications to data storage. There are also several excellent collections of articles on special areas of the subject, such as [17,133], and [128]. The book [15] by Berstel and Perrin treats the subject of variable length codes and contains many ideas related to symbolic dynamics. Finally, there is an excellent recent survey by Boyle on Open Problems in Symbolic Dynamics [24].

## Origins of Symbolic Dynamics: Modeling of Dynamical Systems

Symbolic dynamics began as an effort to model dynamical systems using sequences of symbols. A *dynamical system* is a pair  $(X, T)$  where  $X$  is a set and  $T$  is a transformation from  $X$  to itself. For definiteness, we assume that  $T$  is invertible, although this is not a necessary restriction.



Since  $T$  maps  $X$  to itself, we can iterate  $T$ :  $T^2 = T \circ T$ ,  $T^3 = T \circ T \circ T$ , etc. The *orbit* of a point  $x \in X$  is the sequence of points:  $\dots, T^{-2}(x), T^{-1}(x), x, T(x), T^2(x), \dots$ . In the theory of dynamical systems, one asks questions about orbits such as the following: Are there periodic orbits (i. e.,  $x$  such that  $T^n(x) = x$  for some  $n > 0$ )? Are there dense orbits (the orbit of  $x$  is dense if for any point  $y$  in  $X$ ,  $T^n(x)$  is “close” to  $y$  for some  $n$ )? How does the behavior of an orbit vary with  $x$ ? How can we describe the collection of all orbits of the dynamical system? When is the dynamical system “chaotic”? For more information on dynamical systems, we refer the reader to [16,38,47].

The subject of dynamical systems has its roots in Classical Mechanics; in that setting,  $X$  is the set of all possible states of a system (e. g., the positions, momenta of all particles in a physical system), and the transformation  $T$  is the time evolution map, which maps the state of the system at one time to the state of the system at one time unit later.

Symbolic dynamics provides a model for the orbits of a dynamical system  $(X, T)$  via a space of sequences. This is done by “quantizing”  $X$  into cells, associating symbols to the cells and representing points as bi-infinite sequences of symbols. For instance in Fig. 1,  $X$  is a square, and  $T$  is some transformation of the square.

We have drawn a portion of the orbit of a point  $x \in X$  for the dynamical system  $(X, T)$ . We have also quantized  $X$  into two cells: the left half, called ‘0’, and the right half, called ‘1’. Then the point  $x$  is represented by the bi-infinite sequence  $s(x) = \dots s_{-2}s_{-1}.s_0s_1s_2\dots$  where  $s_n$  is the label of the cell to which  $T^n(x)$  belongs (here, we use the decimal point to separate coordinates  $s_i, i < 0$  from  $s_i, i \geq 0$ ). So, for  $x$  as given in Fig. 1, we see that

$$s(x) = \dots 11.001\dots$$

for instance  $s_0 = 0$  because  $x$  belongs to the left half of the square, and  $s_2 = 1$  because  $T^2(x)$  belongs to the right half

of the square. Now, if  $x$  is represented by the sequence  $s(x)$ , then  $T(x)$  is represented by the shift of  $s(x)$ :

$$s(T(x)) = \dots 110.01\dots$$

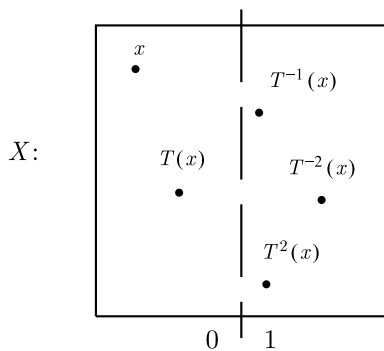
So,  $T$  is represented ‘symbolically’ as the shift transformation.

By representing all points of  $X$  as bi-infinite sequences, we obtain a *symbolic dynamical system*  $(Y, \sigma)$  where  $Y$  is a set of sequences (representing  $X$ ) and  $\sigma$  is the shift transformation (representing  $T$ ). The “symbolic” refers to the symbols, and the “dynamical” refers to the action of the shift transformation.

For this representation to be faithful, distinct points should be represented by distinct sequences, and this imposes extra conditions on how  $X$  is quantized into cells. Also,  $Y$  is typically a set of sequences constrained by certain rules, such as a certain symbol may only be followed by certain other symbols.

In this way, one can use symbolic dynamics to study dynamical systems. Properties of orbits of the original dynamical system are reflected in properties of the resulting sequences. For instance, a point whose orbit is periodic becomes a periodic sequence, and the distribution of the orbit of a point  $x$  in  $X$  is reflected in the distribution of finite strings within  $s(x)$ . Beginning with Hadamard [46] in 1898 and followed by Hedlund, Morse and others in the 1920s, 1930s and 1940s [48,49,92,93], this method was used to prove the existence of periodic, almost periodic and other interesting motions in classical dynamical systems, such as geodesic flows on surfaces of negative curvature; this was done by finding interesting sequences satisfying the constraints defined by the corresponding symbolic dynamical system. Later on, this was extended to general hyperbolic systems, where the symbolic dynamics is constructed using a *Markov Partition*, which is a disjoint collection of open sets whose closures cover  $X$ , each of which looks like a “rectangle”, with vertical (resp., horizontal) fibers contracted (resp., expanded) by  $T$ . Markov Partitions were developed by Adler and Weiss [5], Sinai [120] and Bowen [18,19]; see also [13] and Sect. 6.5 in [78].

In more recent years, symbolic dynamics has been used as a tool in classification problems for dynamical systems. Here, the problem of determining when one dynamical system is ‘equivalent’ to another becomes, via symbolic dynamics, a coding problem. Roughly speaking, two dynamical systems,  $(X_1, T_1)$  and  $(X_2, T_2)$ , are *equivalent* if there is an invertible mapping from  $X_1$  to  $X_2$  which makes  $T_1$  “look like”  $T_2$ . If the corresponding symbolic dynamical systems are denoted  $(Y_1, \sigma)$  and  $(Y_2, \sigma)$ , then an equivalence between  $(X_1, T_1)$  and  $(X_2, T_2)$  becomes a time-in-



**Symbolic Dynamics, Figure 1**  
Representing points symbolically

variant, invertible encoding from  $Y_1$  to  $Y_2$  (time-invariant because the shift transformation represents the dynamics). Thus, the classification problem in dynamical systems leads to a coding problem between constrained sets of sequences.

We have described dynamical systems as the discrete-time iteration of a single mapping. However, continuous-time iterations have been studied since the inception of the subject. These are known as continuous-time flows, with the main example being the set of solutions to a system of ordinary differential equations. Indeed, the work of Hedlund and Morse mentioned above was done in this context.

### Shift Spaces and Sliding Block Codes

Let  $\mathcal{A}$  be an alphabet of symbols, which we assume to be finite. The principal objects of study in symbolic dynamics are certain kinds of collections of sequences of symbols from  $\mathcal{A}$ . Typically, these sequences are infinite  $x = x_0x_1x_2\ldots$ , but it is often more convenient to deal with *bi-infinite sequences*  $x = \ldots x_{-2}x_{-1}x_0x_1x_2\ldots$ . For some problems, the results are similar in the infinite and bi-infinite categories, while for other problems, they are quite different. In this article, we focus on the bi-infinite setting.

The symbol  $x_i$  is the  $i$ th *coordinate* of  $x$ . When writing a specific sequence, we need to specify which is the 0th coordinate. As suggested in Sect. “Origins of Symbolic Dynamics: Modeling of Dynamical Systems”, this is done with a decimal point to separate the  $x_i$  with  $i \geq 0$  from those with  $i < 0$ :  $x = \ldots x_{-2}x_{-1}.x_0x_1x_2\ldots$ . A *block* or *word* over  $\mathcal{A}$  is a finite sequence of symbols from  $\mathcal{A}$ . A block of length  $N$  is called an  $N$ -*block*. For blocks  $u, v$ , the block  $uv$  is the concatenation of  $u$  and  $v$ , and for a block  $w$ , the concatenation of  $N$  copies of  $w$  is denoted  $w^N$ .

The *full  $\mathcal{A}$ -shift*  $\mathcal{A}^{\mathbb{Z}}$  is the set of all bi-infinite sequences of symbols from  $\mathcal{A}$ . The *full  $r$ -shift* is the full shift over the alphabet  $\{0, 1, \ldots, r-1\}$ . The *shift map*  $\sigma$  on a full shift maps a point  $x$  to the point  $y = \sigma(x)$  whose  $i$ th coordinate is  $y_i = x_{i+1}$ .

The *orbit* of a point in a full shift is its orbit under the shift map. The full shift contains many different types of orbits. For instance, it contains a dense orbit (namely, any sequence which contains every block in the alphabet) and periodic orbits (namely, any sequence which is periodic).

We are interested in sets that can be specified by a list (finite or infinite) of forbidden blocks. Namely, given a collection  $\mathcal{F}$  of “forbidden blocks” over  $\mathcal{A}$ , the subset  $X$  consisting of all sequences in  $\mathcal{A}^{\mathbb{Z}}$ , none of whose subwords belong to  $\mathcal{F}$ , is called a *shift space* (or simply *shift*), and we

write  $X = X_{\mathcal{F}}$ . When a shift space  $X$  is contained in a shift space  $Y$ , we say that  $X$  is a *subshift* of  $Y$ .

**Example 1**  $X$  is the set of all binary sequences with no two 1’s next to each other. Here  $X = X_{\mathcal{F}}$ , where  $\mathcal{F} = \{11\}$ . This shift is called the *golden mean shift*, for reasons that will become apparent later.

**Example 2**  $X$  is the set of all binary sequences so that between any two 1’s there are an even number of 0’s. We can take for  $\mathcal{F}$  the collection

$$\{10^{2n+1}1 : n \geq 0\}.$$

This example is naturally called the *even shift*.

**Example 3**  $X$  is the set of all binary sequences such that between any two successive 1’s number of 0’s is prime. We can take for  $\mathcal{F}$  the collection

$$\{10^n1 : n \text{ is composite}\}.$$

This example is naturally called the *prime shift*.

Alternatively (and equivalently), shift spaces can be defined as closed, shift-invariant subsets of full shifts. Here, “closed” means with respect to a metric, for which two points are close if they agree in a large “central block”; one such metric is  $\rho(x, y) = 2^{-k}$  if  $x \neq y$ , with  $k$  maximal such that  $x_{[-k, k]} = y_{[-k, k]}$  (with the conventions that  $\rho(x, y) = 0$  if  $x = y$  and  $\rho(x, y) = 2$  if  $x_0 \neq y_0$ ).

Let  $X$  be a subset of a full shift, and let  $\mathcal{B}_N(X)$  denote the set of all  $N$ -blocks that occur in elements of  $X$ . The *language* of  $X$  is  $\mathcal{B}(X) = \bigcup_N \mathcal{B}_N(X)$ . It can be shown that the language of a shift space determines the shift space uniquely, and so we can equally well describe a shift space by specifying the “occurring” or “allowed” blocks, rather than the forbidden blocks. For example, the golden mean shift is specified by the language of blocks in which 1’s are isolated.

This establishes a connection with automata theory [6,11], which studies collections of blocks, rather than infinite or bi-infinite sequences. The languages that occur in symbolic dynamics, i. e. as  $\mathcal{B}(X)$  for some shift space  $X$ , are simply those sets  $\mathcal{L}$  of blocks that satisfy two simple properties: every sub-block of an element of  $\mathcal{L}$  belongs to  $\mathcal{L}$ , and every element  $w \in \mathcal{L}$  is extendable to a larger block  $awb \in \mathcal{L}$ , with  $a, b \in \mathcal{A}$ .

Some examples are best described by allowed blocks. One example is the famous *Morse shift*.

**Example 4** Let  $A_0 = 0$  and inductively define blocks  $A_{n+1} = A_n \overline{A_n}$ , where  $\overline{A_n}$  denotes bitwise complement. The shift space whose allowed blocks are the subblocks of the  $A_n$  is called the *Morse shift*.

This shift space has an alternative description as follows. Since each  $A_n$  is a prefix of  $A_{n+1}$ , the sequence of blocks  $A_n$  determines a unique right-infinite sequence  $x^+$  (with each  $A_n$  as a prefix). If we denote  $x^-$  as the left-infinite sequence obtained by writing  $x^+$  backwards, then the Morse shift is the closure of the orbit of the point  $x^- . x^+$ . The sequence  $x^+$  is known as the Prouhet–Thue–Morse (PTM) sequence, since Prouhet and Thue introduced it earlier, but for different purposes.

While it is not immediately obvious, it can be shown that the PTM sequence is not periodic; moreover, the Morse shift is a *minimal shift*, which means that all its orbits are dense [91]. In contrast, while the full, golden mean, even and prime shifts all have dense orbits, each also has a dense set of periodic orbits.

There are two simple, but very important, constructions in symbolic dynamics that construct from a given shift space a new version which in some sense looks deeper into the space at the cost of a larger alphabet and more complex description.

Let  $X$  be a shift space over the alphabet  $\mathcal{A}$ , and  $\mathcal{A}^{(N)} = \mathcal{B}_N(X)$ . We can consider  $\mathcal{A}^{(N)}$  as an alphabet in its own right, and form the full shift  $(\mathcal{A}^{(N)})^{\mathbb{Z}}$ . Define the *Nth higher block code* by

$$(\beta_N(x))_i = x_{[i, i+N-1]}.$$

Then the *Nth higher block shift* or *higher block presentation* of a shift space  $X$  is the image  $X^{[N]} = \beta_N(X)$  in the full shift over  $\mathcal{A}^{(N)}$ .

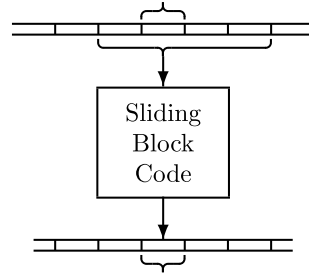
Similarly, define the *Nth higher power code*  $\gamma_N: X \rightarrow (\mathcal{A}^{(N)})^{\mathbb{Z}}$  by

$$(\gamma_N(x))_i = x_{[iN, iN+N-1]}.$$

The *Nth higher power shift*  $X^N$  of  $X$  is the image  $X^N = \gamma_N(X)$  of  $X$ . The difference between  $X^{[N]}$  and  $X^N$  is that the former is constructed by considering overlapping blocks and the latter by non-overlapping blocks.

Next, we turn to mappings between shift spaces. Suppose that  $x = \dots x_{-1}.x_0x_1\dots$  is a sequence in a shift space  $X$  over  $\mathcal{A}$ . We can transform  $x$  into a new sequence  $y = \dots y_{-1}.y_0y_1\dots$  over another alphabet  $\mathcal{C}$  as follows. Fix integers  $m$  and  $n$  with  $-m \leq n$ . To compute the  $i$ th coordinate  $y_i$  of the transformed sequence, we use a function  $\Phi$  that depends on the “window” of coordinates of  $x$  from position  $i-m$  to position  $i+n$ . Here  $\Phi: \mathcal{B}_{m+n+1}(X) \rightarrow \mathcal{C}$  is a fixed *block map*, called an  $(m+n+1)$ -*block map* from allowed  $(m+n+1)$ -blocks in  $X$  to symbols in  $\mathcal{C}$ , and so

$$y_i = \Phi(x_{i-m}x_{i-m+1}\dots x_{i+n}) = \Phi(x_{[i-m, i+n]}).$$



Symbolic Dynamics, Figure 2  
Sliding block code

This is illustrated in Fig. 2.

Let  $X$  be a shift space over  $\mathcal{A}$ , and  $\Phi: \mathcal{B}_{m+n+1}(X) \rightarrow \mathcal{C}$  be a block map. Then the map  $\phi: X \rightarrow \mathcal{C}^{\mathbb{Z}}$  defined by  $y = \phi(x)$ , with  $y_i$  given by  $\Phi$  above, is called the *sliding block code* with *memory*  $m$  and *anticipation*  $n$  induced by  $\Phi$ . We will denote the formation of  $\phi$  from  $\Phi$  by  $\phi = \Phi_{\infty}^{[-m, n]}$ , or more simply by  $\phi = \Phi_{\infty}$  if the memory and anticipation of  $\phi$  are understood. If not specified, the memory is taken to be 0. If  $Y$  is a shift space contained in  $\mathcal{C}^{\mathbb{Z}}$  and  $\phi(X) \subseteq Y$ , we write  $\phi: X \rightarrow Y$ .

In analogy with the characterization of shift spaces as closed shift-invariant sets, sliding block codes can be characterized in a topological manner: namely, as the maps between shift spaces that are continuous and commute with the shift. This result is known as the Curtis–Hedlund–Lyndon theorem [50].

**Example 5** Let  $\mathcal{A} = \{0, 1\} = \mathcal{C}$ ,  $X = \mathcal{A}^{\mathbb{Z}}$ ,  $m = 0$ ,  $n = 1$ , and  $\Phi(a_0a_1) = a_0 + a_1 \pmod{2}$ . Let  $\phi = \Phi_{\infty}: X \rightarrow X$ .

**Example 6** The sliding block code, generated by  $\Phi(00) = 1$ ,  $\Phi(01) = 0 = \Phi(10)$ , maps the golden mean shift onto the even shift.

**Example 7** There is a trivial sliding block code from the full 2-shift into the full 3-shift, generated by  $\Phi(0) = 0$ ,  $\Phi(1) = 1$ .

If a sliding block code  $\phi: X \rightarrow Y$  is onto, then  $\phi$  is called a *factor code* or *factor map*, and  $Y$  is a *factor* of  $X$ . If  $\phi: X \rightarrow Y$  is one-to-one, then  $\phi$  is called an *embedding* of  $X$  into  $Y$ . The sliding block code in Example 7 is an embedding but not a factor code, while the codes in Examples 5 and 6 are factor maps, but not embeddings.

A major (and unrealistic) goal of symbolic dynamics is to classify in an explicit way shift spaces up to the following natural notion of equivalence. A sliding block code  $\phi: X \rightarrow Y$  is a *conjugacy* (or *topological conjugacy*) if it is invertible with sliding block inverse. Equivalently, a conjugacy is a bijective sliding block code and therefore simulta-

neously a factor code and an embedding. If there is a conjugacy from one shift space  $X$  to another  $Y$ , we say that  $X$  and  $Y$  are *conjugate*, denoted  $X \cong Y$ .

As an example, the higher block map  $\beta_N$  is a conjugacy between a shift space  $X$  and its higher block shift  $X^{[N]}$ . Via this code, we can “re-code” any sliding block code as a 1-block code (though typically a conjugacy and its inverse cannot, by this artifice, be simultaneously re-coded to 1-block codes).

In this section, we have given examples of relatively simple sliding block codes. But the typical conjugacy, as well as factor code and embedding, can be much more complicated.

### Shifts of Finite Type and Sofic Shifts

A *shift of finite type* (SFT) is a shift space that can be described by a finite set of forbidden blocks, i. e., a shift space  $X$  having the form  $X_{\mathcal{F}}$  for some finite set  $\mathcal{F}$  of blocks. The terminology shift of finite type (or *subshift of finite type*) comes from dynamical systems (Smale [121]).

An SFT is *M-step* (or has *memory M*) if it can be described by a collection of forbidden blocks all of which have length  $M + 1$ . It is easy to see that any SFT is *M-step* for some  $M$ .

Since any shift space can be defined by many different collections of forbidden blocks, it is useful to have the following equivalent condition expressed in terms of allowed blocks: an SFT is *M-step* if and only if whenever  $u$  is an allowed block of length at least  $M$ ,  $u'$  is the suffix of  $u$  with length  $M$  and  $a$  is a symbol, then  $ua$  is allowed if and only if  $u'a$  is allowed. In other words, in order to tell whether a symbol can be allowably concatenated to the end of an allowed word  $u$ , one need only look at the last  $M$  symbols of  $u$ . This is analogous to the “finite memory” property of *M-step* Markov chains.

The golden mean shift  $X$  is a 1-step SFT, since it was defined by a forbidden list consisting of exactly one block:  $\mathcal{F} = \{11\}$ . Equivalently, it is only the last symbol of an allowed block that determines whether a given symbol can be concatenated at the end. In contrast, the even shift is not an SFT: for any  $M$ , the symbol 1 can be concatenated to the end of exactly one of the (allowed) words  $10^M$  and  $10^{M+1}$ .

Recall that the higher block code  $\beta_M$  is a conjugacy from  $X$  to  $X^{[M]}$ . Via this code, any SFT can be recoded to a 1-step SFT. And so any sliding block code on a shift space can be recoded to a 1-block code on a 1-step SFT. It is useful to have a concrete description of 1-step SFT's. In fact, these are precisely the shift spaces consisting of all bi-infinite sequences of vertices along paths on a finite di-

rected graph. These are called *vertex shifts*. We find it more convenient to work with sequences of edges instead. To be precise:

Let  $G$  be a finite directed graph (or simply graph) with vertices (or *states*)  $\mathcal{V} = \mathcal{V}(G)$  and *edges*  $\mathcal{E} = \mathcal{E}(G)$ . For an edge  $e$ ,  $i(e)$  denotes the initial state and  $t(e)$  the terminal state. A *path* in  $G$  is a finite sequence of edges in  $G$  such that the terminal state of an edge coincides with the initial state of the following edge; a *cycle* in  $G$  is path that begins and ends at the same state. We will assume that  $G$  is *essential*, i. e., that every state has at least one outgoing edge and one incoming edge.

The *adjacency matrix*  $A = A(G)$  is the matrix indexed by  $\mathcal{V}$  with  $A_{IJ}$  equal to the the number of edges in  $G$  with initial state  $I$  and terminal state  $J$ . Since a graph and its adjacency matrix essentially determine the same information, we will frequently associate a graph  $G$  with its adjacency matrix  $A$  and a nonnegative integer matrix  $A$  with a graph  $G$ .

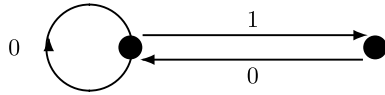
The *edge shift*  $X_G$  or  $X_A$  is the shift space over the alphabet  $\mathcal{A} = \mathcal{E}$  defined by

$$\begin{aligned} X_G &= X_A \\ &= \{ \xi = (\xi_i)_{i \in \mathbb{Z}} \in \mathcal{E}^{\mathbb{Z}} : \text{each } \xi_{i+1} \text{ follows } \xi_i \} . \end{aligned}$$

It can be readily verified that edge shifts are 1-step SFT's. While edge shifts do not include all 1-step SFT's, any 1-step SFT can be recoded to an edge shift, and, compared with vertex shifts, edge shifts offer the advantage of a more compact description.

For many purposes, one can study a general shift space  $X$  by breaking it into smaller, more well-behaved pieces. A shift space is *irreducible* if whenever  $u$  and  $w$  are allowed blocks, there is a “connecting” block  $v$  such that  $uvw$  is allowed. While shift spaces do not always decompose into disjoint unions of irreducible shifts, every SFT can be written as a finite disjoint union of irreducible SFT's  $X_i$  together with “transient” one-way connections from one  $X_i$  to another. And irreducible edge shifts can be characterized in a particularly concrete form: namely,  $X_G$  is irreducible if and only if  $G$  is *irreducible*, i. e., for every ordered pair of vertices  $I$  and  $J$  there is a path in  $G$  starting at  $I$  and ending at  $J$ .

There is a stronger notion which is defined by a uniformity condition on the length of the connecting block. A shift space is *mixing* if whenever  $u$  and  $w$  are allowed blocks, there is an  $N$ , possibly depending on  $u$  and  $w$ , such that for all  $n \geq N$ , there is block  $v$  of length  $n$  such that  $uvw$  is allowed. And an edge shift  $X_G$  is mixing if and only if  $G$  is *primitive*, i. e., there is an integer  $N$  such that for any  $n \geq N$  and any ordered pair of vertices  $I$  and  $J$  there is



Symbolic Dynamics, Figure 3  
Presentation of golden mean shift



Symbolic Dynamics, Figure 4  
Presentation of even shift

a path in  $G$  of length  $n$  starting at  $I$  and terminating at  $J$ . It follows that for SFT's in the definition of mixing, the uniform connecting length  $N$  can be chosen independent of the allowed blocks  $u$  and  $w$ .

It can be shown that, in some sense, any irreducible SFT  $X$  can be broken down into a union of disjoint maximal mixing shifts; namely,  $X$  can be written as the disjoint union of finitely many sets  $X_i$ ,  $i = 0, \dots, p-1$  such that  $\sigma(X_i) = X_{i+1} \bmod p$  and for each  $i$ ,  $\sigma^p$  restricted to  $X_i$  can be regarded as a mixing SFT. This is a consequence of Perron–Frobenius theory, upon which symbolic dynamics relies heavily; see Seneta [118] for an introduction to this theory.

A *sofic shift* is the set of bi-infinite sequences obtained from a finite labeled directed graph  $G = (G, \mathcal{L})$ ; here,  $G$  is a finite directed graph and  $\mathcal{L}$  is a labeling of the edges of  $G$ . The labeled graph is often called a *presentation* of the sofic shift. The golden mean shift and even shift are sofic, with presentations given in Figs. 3 and 4.

SFT's are sofic because any  $M$ -step SFT can be presented by a graph whose states are allowed  $M$ -blocks. Note also that any sofic shift is a factor of an SFT, namely via a (1-block) factor code  $\mathcal{L}_\infty$  on the edge shift  $X_G$  based on a presentation  $(G, \mathcal{L})$ . In fact, the converse is true, and so the sofic shifts are precisely the shift spaces that are factors of SFT's. This was the original definition of sofic shifts given by Weiss [130].

Typically, the labeling is *right resolving*, which means that at any given state, all outgoing edges have distinct labels (as in Figs. 3 and 4).

### Theorem 1

- (a) Any sofic shift has a right resolving presentation.
- (b) Any irreducible sofic shift has a unique minimal right resolving presentation.

Part (a) is a direct consequence of the subset construction in automata theory [6,11] which constructs a right resolv-

ing presentation from an arbitrary presentation; see also Coven and Paul [33,34]. Part (b) makes use of the state-minimization algorithm from automata theory [6,11], but requires an idea beyond that found in automata theory (Fischer [39,40]). The unique presentation in part (b) may be regarded as a canonical presentation. We remark that for an irreducible (resp. mixing) sofic shift, the underlying graph of the unique minimal right resolving presentation is irreducible (resp., primitive).

Sometimes it is useful to weaken the concept of right resolving to *right closing*, which means “right resolving with delay”; more precisely a labeling is right closing, with delay  $D$  if all paths of length  $D+1$  with the same initial state and the same label have the same initial edge. Also, we sometimes consider *left resolving* and *left closing* labellings (replace “outgoing” with “incoming” in the definition of right resolving, and replace “initial” with “terminal” in the definition of right closing).

While the class of SFT's is defined by a “finite-memory” property, the more general class of sofic shifts is defined by a “finite-state” property, in that the possible symbols that can occur at time 0 are determined by the past via one of finitely many states. In a presentation, the vertices can be viewed as state information which connects sequences in the past with sequences in the future.

It is not difficult to show that neither the prime shift (Example 3) nor the Morse shift (Example 4) are sofic and therefore also not SFT. For the prime shift, this can be done by an application of the pumping lemma from automata theory [6] (or p. 68 of [78]).

There are uncountably many shift spaces, but only countably many sofic shifts. So, it is not surprising that the behavior of sofic shifts is very special. However, they are very useful in modeling smooth dynamical systems (Sect. “Origins of Symbolic Dynamics: Modeling of Dynamical Systems”) and in information theory and applications to data recording (Sect. “Coding for Data Recording Channels”), where they arise as *constrained systems*, although this term is usually reserved for the set of (finite) blocks obtained from a finite directed labeled graph [83].

### Entropy and Periodic Points

An invariant of conjugacy is an object associated to a shift space that is preserved under conjugacy. It can be shown that many of the concepts that we have already introduced are invariants: irreducibility, mixing, as well as the properties of being a shift of finite type or sofic shift. Beyond these qualitative invariants, there are many quantitative invariants that can, in many cases, be computed explicitly. Foremost among these is topological entropy.



The (topological) entropy (or simply entropy) of  $X$  is:

$$h(X) = \lim_{N \rightarrow \infty} \frac{\log |\mathcal{B}_N(X)|}{N};$$

here,  $|\cdot|$  denotes cardinality, and for definiteness the log is taken to mean  $\log_2$  but any base will do. A subadditivity argument shows that the limit does indeed exist [127]. It should be evident that  $h(X)$  is a measure of the “size” or “complexity” of  $X$ , as it is simply the asymptotic growth rate of the number of blocks that occur in  $X$ . Topological entropy for continuous dynamical systems was defined by Adler, Konheim and McAndrew [3], in analogy with measure-theoretic entropy.

Since a  $k$ -block sliding block code from a shift space  $X$  to a shift space  $Y$  maps  $\mathcal{B}_{N+k-1}(X)$  into  $\mathcal{B}_N(Y)$ , it is not hard to see that entropy is an invariant of conjugacy and that it cannot increase under factors and cannot decrease under embeddings.

In many cases, one can explicitly compute entropy. For example, for the full  $r$ -shift  $X$ ,  $|\mathcal{B}_N(X)| = r^N$ , and so  $h(X) = \log r$ . And from the defining sequence of the Morse shift, we see that the number of distinct  $2^N$ -blocks is at most  $4(2^N)$ , it follows that the growth rate cannot be exponential and so the entropy of the Morse shift is zero.

For SFT's and more generally for sofic shifts, entropy can be computed explicitly. The key to this computation is the following result, which is based on the Perron–Frobenius theorem.

**Theorem 2** [99,119] *For any graph  $G$ ,  $h(X_G) = \log \lambda_{A(G)}$ , where  $\lambda_{A(G)}$  is the largest eigenvalue of  $A(G)$ .*

The rough idea is that the number of  $N$ -blocks in  $X_G$  is the number of paths of length  $N$  and thus also the sum of the entries of  $A(G)^N$ , which is controlled by the largest eigenvalue. This is proven first for primitive graphs, whose adjacency matrices have a unique eigenvalue of maximum modulus and this eigenvalue is positive; in fact, for a primitive graph and a pair of states  $I, J$ , the number of paths of length  $N$  from  $I$  to  $J$  grows like  $\lambda_{A(G)}^N = 2^{Nh(X_G)}$ . For general graphs, one uses the decomposition into primitive, and then irreducible, graphs.

To extend this to a sofic shift  $Y$ , one uses a right resolving presentation  $(G, \mathcal{L})$  of  $Y$ . Since every block of  $Y$  is the label of at most  $V(G)$  paths in  $G$ , it follows that:

**Theorem 3** *Let  $\mathcal{G} = (G, \mathcal{L})$  be a right-resolving labeled graph presenting a sofic shift  $Y$ . Then  $h(Y) = h(X_G)$ .*

From Fig. 3, we see that the golden mean shift is obtained as a right resolving presentation of the graph with adjac-

ency matrix:

$$A = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}.$$

A computation shows that  $\lambda_A$  is the golden mean, and so the entropy of the golden mean shift is the log of the golden mean; this is one explanation of the meaning of the term golden mean shift. From Fig. 4, we see that the even shift has the same entropy.

One cannot overstate the importance of entropy as an invariant. Yet, it is somewhat crude; it is perhaps not surprising that a single numerical invariant would not be sufficient to completely capture the many intricacies of shift spaces, even of sofic shifts or SFT's.

An invariant finer than entropy is the zeta function, which combines information regarding the numbers of periodic sequences of all periods, described as follows.

For a shift space  $X$ , let  $p_n(X)$  denote the number of points in  $X$  of period  $n$  (i.e., the number of  $x \in X$  such that  $\sigma^n(x) = x$ ). It is straightforward to show that each  $p_n(X)$  is an invariant. Since distinct periodic sequences define distinct blocks, it follows that

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log p_n(X) \leq h(X).$$

The inequality can be strict; for example, there are shift spaces (such as the direct product of the Morse shift with the full 2-shift) with positive entropy but no periodic points at all.

However, for irreducible SFT's and sofic shifts, the entropy  $h(X)$  can be recovered from the sequence  $p_n(X)$ . The key to understanding this is the fact that for an edge shift  $X = X_A$ ,  $p_n(X) = \text{tr}(A^n)$  and thus is the sum of the  $n$ th powers of the (non-zero) eigenvalues of  $A$ ; in the case that  $A$  is primitive, the largest eigenvalue  $\lambda_A$  strictly dominates the other eigenvalues, and thus for large  $n$ ,

$$\log p_n(X) = \log \text{tr}(A^n) \sim n \log \lambda_A \sim nh(X).$$

This shows that for a mixing SFT, the entropy equals the growth rate of numbers of periodic points. In fact, this result applies to all SFT's and sofic shifts.

**Theorem 4** *For a sofic shift  $X$ ,*

$$\limsup_{n \rightarrow \infty} \frac{1}{n} \log p_n(X) = h(X).$$

The lim sup turns out to be a limit in the case that  $X$  is a mixing sofic shift.

Most of what we have stated for  $p_n(X)$  here applies equally well to  $q_n(X)$ , the number of points of *least* period  $n$  in  $X$ . This follows from the fact that “most” periodic points of period  $n$  have least period  $n$ .

The periodic point information can be conveniently combined into a single invariant, known as the *zeta function*. For a shift space  $X$ ,

$$\zeta_X(t) = \exp \left( \sum_{n=1}^{\infty} \frac{p_n(X)}{n} t^n \right).$$

For an edge shift  $X_A$ , one computes the zeta function to be the reciprocal of a polynomial:

$$\zeta_{X_A}(t) = \frac{1}{t^r \chi_A(t^{-1})} = \frac{1}{\det(I - tA)},$$

which is completely determined by the non-zero eigenvalues (with multiplicity) of  $A$  (Bowen and Lanford [21]). As an example, the zeta function of the golden mean shift is  $1/(1 - t - t^2)$ .

The discussion above applies equally well to SFT's since they are conjugate to edge shifts. For a sofic shift, the zeta function turns out to be a rational function, i. e., quotient of two polynomials. This can be shown by analyzing properties of a right resolving presentation of the sofic shift. From this it turns out that the zeta function of the even shift is  $(1 - t)/(1 - t - t^2)$ . The technique for computing zeta functions of sofic shifts was developed by Manning [80] (actually, Manning developed the technique to compute zeta functions of hyperbolic dynamical systems).

So, for sofic shifts, all of the periodic point information is determined by a finite collection of complex numbers, namely the zeros and poles of the zeta function.

Finally, we mention another simple invariant obtained from the periodic points. The *period*,  $\text{per}(X)$ , of a shift space  $X$  is the gcd of lengths of periodic points in  $X$ , i. e., the gcd of the set of  $n$  such that  $p_n(X) \neq 0$ . If  $X = X_G$  is an edge shift and  $G$  is irreducible, then  $\text{per}(X_G) = \text{per}(G)$ , which is defined to be the gcd of cycle lengths in  $G$  and coincides with the gcd of the lengths of all cycles in  $G$  based at any given state. For an irreducible graph with period  $p$  and any state  $I$ , the number of cycles of length  $N$ , a multiple of  $p$ , at  $I$  grows like  $\lambda_{A(G)}^N = 2^{N h(X_G)}$ . Also, an irreducible graph  $G$  is primitive if  $\text{per}(G) = 1$ .

## The Conjugacy Problem

The conjugacy problem for SFT's and sofic shifts is a major open problem. After much effort, it remains unsolved today. Much of what we know goes back to R. Williams [131] in the 1970s. One of Williams' main results was that any

conjugacy can be decomposed into simple building blocks, as follows.

Let  $A$  and  $B$  be nonnegative integral matrices, with associated graphs  $G$  and  $H$ . An *elementary equivalence from  $A$  to  $B$*  is a pair  $(R, S)$  of rectangular nonnegative integral matrices satisfying

$$A = RS, \quad B = SR. \quad (1)$$

In this case we write  $(R, S): A \equiv B$ . A *strong shift equivalence of lag  $\ell$  from  $A$  to  $B$*  is a sequence of  $\ell$  elementary equivalences

$$(R_1, S_1): A = A_0 \equiv A_1,$$

$$(R_2, S_2): A_1 \equiv A_2,$$

$$\dots,$$

$$(R_\ell, S_\ell): A_{\ell-1} \equiv A_\ell = B.$$

In this case we write  $A \approx B(\text{lag } \ell)$ . We say that  $A$  is *strong shift equivalent to  $B$*  (and write  $A \approx B$ ) if there is a strong shift equivalence of some lag from  $A$  to  $B$ .

Via the matrix Equation (1), one creates a graph  $K$  whose state set is the disjoint union of  $\mathcal{V}(G)$  and  $\mathcal{V}(H)$ ; for each  $I \in V_G$  and  $J \in V_H$ , the graph has  $R_{IJ}$  edges from  $I$  to  $J$  and  $S_{JI}$  edges from  $J$  to  $I$ . The equation  $A = RS$  allows one to associate each edge  $e$  of  $G$  with a unique path  $r(e)s(e)$  of length two running from  $\mathcal{V}(G)$  to  $\mathcal{V}(H)$  to  $\mathcal{V}(G)$ . Similarly, the equation  $B = SR$  allows one to associate each edge  $e$  of  $H$  with a path  $s(e)r(e)$  of length two running from  $\mathcal{V}(H)$  to  $\mathcal{V}(G)$  to  $\mathcal{V}(H)$ . One can show that the 2-block sliding block code defined by  $\Phi(e f) = s(e) r(f)$  defines a conjugacy from  $X_A$  to  $X_B$ , and so whenever  $A \approx B$ ,  $X_A \cong X_B$ . Williams proved this and its converse:

**Theorem 5** (R. Williams [131]) *The edge shifts  $X_A$  and  $X_B$  are conjugate if and only if  $A$  and  $B$  are strong shift equivalent.*

In fact, Williams showed that any conjugacy can be decomposed into a composition of conjugacies defined by elementary equivalences. This can be interpreted in a way that shows that  $X_A$  and  $X_B$  are conjugate if and only if we can pass from  $G$  to  $H$  by a sequence of state splittings and amalgamations, defined as follows.

Let  $G$  be a graph with states  $\mathcal{V}$  and edges  $\mathcal{E}$ . For each  $I \in \mathcal{V}$ , partition the outgoing edges from  $I$  into disjoint nonempty sets  $\mathcal{E}_I^1, \mathcal{E}_I^2, \dots, \mathcal{E}_I^{m(I)}$ . Let  $\mathcal{P}$  denote the resulting partition of  $\mathcal{E}$ , and let  $\mathcal{P}_I$  denote the partition  $\mathcal{P}$  restricted to  $\mathcal{E}_I$ . The *out-split graph  $H$  formed from  $G$  using  $\mathcal{P}$*  has states  $I^1, I^2, \dots, I^{m(I)}$ , where  $I$  ranges over the states in  $\mathcal{V}$ , and edges  $e^i$ , where  $e$  is any edge in  $\mathcal{E}$  and

$1 \leq j \leq m(t(e))$ . If  $e \in \mathcal{E}$  goes from  $I$  to  $J$ , then  $e \in \mathcal{E}_I^j$  for some  $i$ , and we define the  $e^j$  to have initial state  $I^i$  and terminal state  $J^j$ . The 2-block code generated by  $\Phi(e^j) = e^j$ , where  $f \in \mathcal{E}_{t(e)}^j$ , defines a conjugacy, called an out-splitting code, from  $X_G$  to  $X_H$ . Similarly, one defines an in-splitting code. Inverses of these conjugacies are called out-amalgamation and in-amalgamation codes.

Figure 5 depicts an out-splitting. The graph (a) on the left has three states  $I, J, K$  and the partition elements that define the splitting are  $\mathcal{E}_I^1 = \{a\}$ ,  $\mathcal{E}_I^2 = \{b, c\}$ ,  $\mathcal{E}_J^1 = \{d\}$ ,  $\mathcal{E}_K^1 = \{e\}$ ,  $\mathcal{E}_K^2 = \{f\}$ ; the graph (b) is the resulting split graph.

By interpreting state splitting and amalgamation in terms of adjacency matrices one can show that such operations generate elementary equivalences. And one can decompose elementary equivalences into splittings and amalgamations. It follows that:

**Theorem 6** (R. Williams [131]) *Every conjugacy from one edge shift to another is the composition of splitting codes and amalgamation codes.*

This classification for edge shifts naturally extends to SFT's since every SFT is conjugate to an edge shift. It also extends to sofic shifts, and we describe this in the context of irreducible sofic shifts.

Recall that an irreducible sofic shift has a unique minimal right resolving presentation. Any labeled graph can be completely described by a symbolic adjacency matrix, which records the transitions (edges) in the underlying physical graph, as well as the labels of the edges. Namely, the symbolic adjacency matrix is indexed by the states of the underlying graph and the  $(I, J)$ -entry is the formal sum of the labels of edges from  $I$  to  $J$ . It turns out that the notions of elementary equivalence, and hence strong shift equivalence, can be extended to more general categories, in particular to symbolic adjacency matrices.

**Theorem 7** (Krieger [72], Nasu [97]) *Let  $X$  and  $Y$  be irreducible sofic shifts. Let  $A$  and  $B$  be the symbolic adjacency matrices of the minimal right-resolving presentations of  $X$  and  $Y$ , respectively. Then  $X$  and  $Y$  are conjugate if and only if  $A$  and  $B$  are strong shift equivalent.*

The classification, provided by these results, would be of limited use if the story ended here. Fortunately, Williams showed that strong shift equivalence yields a strong, delicate and somewhat computable necessary condition for conjugacy.

Let  $A$  and  $B$  be nonnegative integral matrices and  $\ell \geq 1$ . A *shift equivalence* of lag  $\ell$  is a pair  $(R, S)$  of rectangular nonnegative integral matrices satisfying the shift

equivalence equations

$$AR = RB, \quad SA = BS, \quad A^\ell = RS, \quad B^\ell = SR.$$

We denote this situation by  $(R, S): A \sim B(\text{lag } \ell)$ . We say that  $A$  is *shift equivalent* to  $B$ , written  $A \sim B$ , if there is a shift equivalence from  $A$  to  $B$  of some lag. It is not hard to see that an elementary equivalence is a shift equivalence of lag 1 and that shift equivalence is an equivalence relation. It follows that:

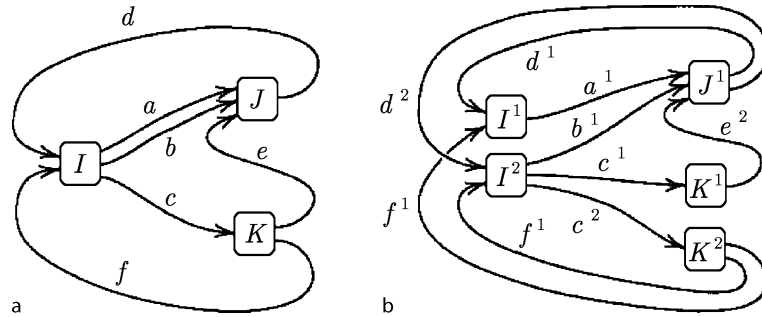
**Theorem 8** (Williams [131]) *Strong shift equivalence implies shift equivalence. More precisely, if  $A \approx B(\text{lag } \ell)$ , then  $A \sim B(\text{lag } \ell)$ .*

Recall from Sect. "Entropy and Periodic Points" that for an edge shift  $X_A$ , the set of nonzero eigenvalues, with multiplicity, of  $A$  determines the zeta function and hence this set is an invariant of conjugacy. Using the shift equivalence equations, one can show more: the entire Jordan form corresponding to the nonzero eigenvalues is an invariant. This information depends only on properties of the adjacency matrix considered as a linear transformation (over  $\mathbf{R}$  or  $\mathbf{Q}$ ). However,  $A$  is a nonnegative, integral matrix and both nonnegativity and integrality provide substantially more information. One such invariant, that follows from shift equivalence and makes use of integrality is the Bowen–Franks group [20],  $BF(A) = Z'/Z'(I - A)$ .

Until recently all of the information contained in known conjugacy invariants, such as those above, was subsumed in shift equivalence. And Kim and Roush showed that shift equivalence is decidable [60,61], meaning that there is a finite decision procedure via a Turing machine that decides whether two given edge shifts, and therefore two given SFT's, are conjugate (they also showed that a notion of shift equivalence for sofic shifts, formulated by Boyle and Krieger [26], is decidable [62]). So, a central focus of the subject was the question: is shift equivalence a complete invariant of conjugacy? The answer turns out to be No, as proven by Kim and Roush [63]; see also the survey article [125]. However, it is a complete invariant of a weaker form of conjugacy; we say that  $X_A$  and  $X_B$  are *eventually conjugate* if all sufficiently large powers are conjugate. It is not hard to show that if  $A \sim B$ , then  $X_A$  and  $X_B$  are eventually conjugate, and the converse is true as well:

**Theorem 9** (Kim and Roush [60], Williams [131]) *Edge shifts  $X_A$  and  $X_B$  are eventually conjugate if and only if  $A$  and  $B$  are shift equivalent.*

Also the Kim–Roush counterexamples and subsequent work do not bear on a special case of the conjugacy problem: if  $A$  is shift equivalent to the  $1 \times 1$  matrix  $[n]$ , is



Symbolic Dynamics, Figure 5

A state splitting

$X_A$  conjugate to the full  $n$ -shift? This question, known as the *little shift equivalence problem*, is particularly intriguing because in this case the condition  $A \sim [n]$  is simply the statement that  $A$  has exactly one non-zero eigenvalue, namely  $n$ .

Shift equivalence can be characterized in another way that has turned out to be very useful. Let  $A$  be an  $r \times r$  integral matrix. Let  $R_A$  denote the real eventual range of  $A$ , i.e.,  $R_A = R^r A^r$ . The *dimension group* of  $A$  is defined:

$$\Delta_A = \{v \in R_A : vA^k \in \mathbb{Z}^r \text{ for some } k \geq 0\}.$$

The *dimension group automorphism*  $\delta_A$  of  $A$  is the restriction of  $A$  to  $\Delta_A$ , so that  $\delta_A(v) = vA$  for  $v \in \Delta_A$ . The *dimension pair* of  $A$  is  $(\Delta_A, \delta_A)$ .

If  $A$  is also nonnegative, then we define the *dimension semigroup* of  $A$  to be

$$\Delta_A^+ = \{v \in R_A : vA^k \in (\mathbb{Z}^+)^r \text{ for some } k \geq 0\}.$$

The *dimension triple* of  $A$  is  $(\Delta_A, \Delta_A^+, \delta_A)$ .

It can be shown that the dimension triple completely characterizes shift equivalence, i.e., two nonnegative integral matrices are shift equivalent if and only if their dimension groups are isomorphic by an isomorphism that preserves the dimension semigroup and intertwines the dimension group automorphisms. Also, by associating equivalence classes of certain subsets of the shift space  $X_A$  to elements of  $\Delta_A^+$ , one can interpret the dimension triple in terms of the action of the shift map on  $X_A$  [27,68]. And the dimension triple arises prominently in the study of the automorphism group of an SFT, which we now briefly describe. The dimension group for SFT's was developed by Krieger [68,69].

In many areas of mathematics, objects are studied by means of their symmetries. This holds true in symbolic dynamics, where symmetries are expressed by automorphisms. An *automorphism* of a shift space  $X$  is a conjugacy

from  $X$  to itself. The set of all automorphisms of a shift space  $X$  is a group under composition, and is naturally called the *automorphism group*, denoted  $\text{aut}(X)$ .

The goals are to understand  $\text{aut}(X)$  as a group (What kinds of subgroups does it contain? How “big” is it?) and how it acts on  $X$ , e.g., given shift-invariant subsets  $U, V$ , such as finite sets of periodic points, when is there an automorphism of  $X$  that maps  $U$  to  $V$ ? One might hope that the automorphism group would shed new light on the conjugacy problem for SFT's. Indeed, tools developed to study the automorphism group eventually paved the way for Kim and Roush to find examples of shift equivalent matrices that are not strong shift equivalent. On the other hand, the automorphism group cannot tell the entire story. For instance,  $\text{aut}(X_A)$  and  $\text{aut}(X_{A^\top})$  are isomorphic, since any automorphism read backwards can be viewed as an automorphism of the transposed shift, yet  $X_A$  and  $X_{A^\top}$  may fail to be conjugate (for an example due to Kollmer, see p. 81 of [105]). It is not even known if the automorphism groups of the full 2-shift and the full 3-shift are isomorphic.

A good deal of our understanding of the action of the automorphism group on an SFT comes from understanding its induced representation as an action of the dimension group; this action is known as the *dimension representation*. For a much more thorough exposition on  $\text{aut}(X)$ , we refer the reader to Wagoner [125,126].

## Other Coding Problems

The difficulties encountered in attempts to solve the conjugacy problem motivated the formulation and study of weaker, but meaningful, notions of equivalence. For instance, we might say that two shift spaces are equivalent if one can be invertibly encoded to the other by some kind of “finite-state machine”. A precise version of this is as follows.



Shift spaces  $X$  and  $Y$  are *finitely equivalent* if there is an SFT  $W$  together with finite-to-one factor codes  $\phi_X: W \rightarrow X$  and  $\phi_Y: W \rightarrow Y$ . We call  $W$  a *common extension*, and  $\phi_X, \phi_Y$  the *legs*.

Here, by “finite-to-one” we mean merely that each point has a finite number of inverse images. It can be shown that any finite-to-one factor code from one shift space to another must preserve entropy, and so entropy is an invariant of finite equivalence. For irreducible sofic shifts, entropy is a complete invariant:

**Theorem 10** (Parry [100]) *Two irreducible sofic shifts are finitely equivalent if and only if they have the same entropy.*

Note that from this result and the fact that finite-to-one codes between general shift spaces preserve entropy, for irreducible sofic shifts, we could have just as well defined finite equivalence with the common extension  $W$  merely being a shift space. However, if  $W$  is an SFT, we get a more concrete coding interpretation as follows. First, we recode  $W$  to an edge shift  $X_G$  and recode the legs,  $\phi_X = (\Phi_X)_\infty$  and  $\phi_Y = (\Phi_Y)_\infty$ , to one-block codes, and (with a bit more argument) we can assume that  $G$  is irreducible. In this set-up, the finite-to-one condition translates to the so-called “no-diamond” condition, which means that for any given pair of states  $I, J$  and finite sequence  $w$ , there is at most one path from  $I$  to  $J$  with label  $w$  [33,34]. Since, for any fixed state  $I$ , the number of cycles of length  $n$ , a multiple of  $p = \text{per}(G)$ , at  $I$  grows like  $2^{nh(W)}$ , we have, in this set-up a means to invertibly encode a “large” set of allowed blocks in  $X$  to allowed blocks in  $Y$ : namely, fix state  $I$ , and a large  $n$ , which is a multiple of  $p$ ; for any cycle  $\gamma$  of length  $n$  at state  $I$  encode the  $\Phi_X$ -label of  $\gamma$  to the  $\Phi_Y$ -label of  $\gamma$ .

For encoding and decoding, one can dispense with state information if the legs are “almost invertible”. A factor code  $\phi$  is *almost invertible* if it is one-to-one on sequences that are typical in the following sense:  $x$  is typical if every allowed block appears infinitely often in  $x$  both to the left and the right. We then say that shift spaces  $X$  and  $Y$  are *almost conjugate* if there is an SFT  $W$  and almost invertible factor codes  $\phi_X: W \rightarrow X$ ,  $\phi_Y: W \rightarrow Y$ . We call  $(W, \phi_X, \phi_Y)$  an *almost conjugacy* between  $X$  and  $Y$ .

For irreducible sofic shifts, it can be shown that any almost invertible factor code is finite-to-one, and so almost conjugacy implies finite equivalence. Thus, entropy is again an invariant, and together with a second very mild invariant, it is complete:

**Theorem 11** (Adler–Marcus [4]) *Let  $X$  and  $Y$  be irreducible sofic shifts with minimal right resolving presentations  $(G, \mathcal{L})$  and  $(H, \mathcal{M})$ . Then  $X$  and  $Y$  are almost conjugate if and only if  $h(X) = h(Y)$  and  $\text{per}(G) = \text{per}(H)$ .*

In particular, if  $X$  and  $Y$  are mixing, then  $\text{per}(G) = 1 = \text{per}(H)$  and entropy itself is a complete invariant.

Thus, with an almost conjugacy, one can invertibly encode most sequences in  $X$  to those of  $Y$  without the need for auxiliary state information. Moreover, if  $X$  and  $Y$  are almost conjugate, then there is an almost conjugacy of  $X$  and  $Y$  in which one leg is right-resolving and the other leg is left-resolving (and the common extension is irreducible [4]). This gives an even more concrete interpretation to the encoding.

The proofs of Theorems 10 and 11 are actually quite constructive. For illustration we consider a very special, but historically important, case.

Let  $G$  be a graph with constant out-degree  $n$ . A *road coloring*  $\Phi$  is a labeling of  $G$  such that at each state of  $G$ , each symbol  $0, \dots, n-1$  appears exactly once as the label of an outgoing edge. An  $n$ -ary word  $w$  is *synchronizing* if all paths that are labeled  $w$  end at the same state. Figure 6 gives examples of road-colorings, with  $n = 2$ .

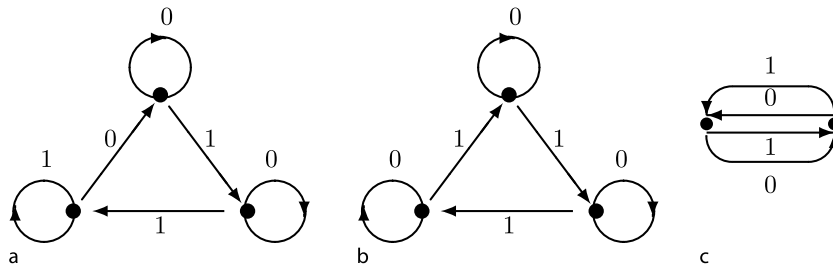
For a road-coloring, a binary word may be viewed as a sequence of instructions given to drivers starting at each of the states. A synchronizing word is a word that drives everybody to the same state. For instance, in Fig. 6a, the word 11 drives everybody to the state in the lower-left corner. But Fig. 6b does not have a synchronizing word because whenever a driver takes a ‘0’ road he stays where he is and whenever a driver takes a ‘1’ road he rotates by 120 degrees. The road-coloring in Fig. 6c is essentially the only road-coloring of its underlying graph, and there is no synchronizing word because drivers must always oscillate between the two states.

Now, let  $X$  be the full  $n$ -shift and  $Y$  be an irreducible SFT with entropy  $\log n$ . Suppose that we could find a presentation  $(G, \mathcal{L})$  of  $Y$  with  $G$  having constant out-degree  $n$  and  $\mathcal{L}_\infty$  finite-to-one. Then, define  $\Phi$  to be any road coloring of  $G$ . Then  $\Phi_\infty$  would be a finite-to-one (in fact, right resolving!) factor code from  $X_G$  to  $X$ . And we would obtain a finite equivalence with  $X = X_G$ ,  $\phi_Y = \mathcal{L}_\infty$  and  $\phi_X = \Phi_\infty$ . If, moreover, we could choose  $\mathcal{L}$  and  $\Phi$  such that  $\phi_Y$  and  $\phi_X$  are almost invertible, then we would have an almost conjugacy.

It turned that this could be arranged for  $\phi_Y$  via a construction related to state splitting [2]. And if  $Y$  were mixing, then  $G$  could be chosen to be primitive and  $\phi_Y$  almost invertible. In this setting,  $\phi_X = \Phi_\infty$  would be almost invertible iff  $\Phi$  has a synchronizing word; the sufficiency follows from the fact that every bi-infinite binary sequence which contains  $w$  infinitely often to the left would be the label of exactly one bi-infinite sequence of edges (to see this, use the synchronizing and road-coloring properties).

The construction of such a labeling  $\Phi$  became known





Symbolic Dynamics, Figure 6  
Some road-colorings

as the Road Problem, which remained open for thirty years. In the meantime, a weaker version of the the road problem was solved and applied to yield this special case of Theorem 11 (see [2]). Nevertheless, the problem remained an important problem in graph/automata theory and was only recently solved:

**Theorem 12** (Road theorem (Trachtman [122])) *If  $G$  is a finite directed primitive graph with constant out-degree  $n$ , there a road-coloring of  $G$  which has a synchronizing word.*

Trachtman's approach relies heavily on earlier work of Friedman [44] and Kari [57].

The primitivity assumption above is close to necessary. Clearly some kind of connectivity is required and in the presence of irreducibility, primitivity would be necessary since otherwise there would be a "phase" introduced in the graph that would never allow a word to synchronize, as in Fig. 6c.

So far, we have focused on equivalences between symbolic systems. There has also been considerable attention paid to problems of embedding one system into another and factoring one onto another.

One of the most striking results of this type is the Krieger embedding theorem. It is not hard to show that any proper subshift of an irreducible SFT must have strictly smaller entropy. Thus, a necessary condition for a proper embedding of a shift space into an irreducible SFT is that it have strictly smaller entropy. This condition, together with a trivially necessary condition on periodic points, turns out to be sufficient. Recall that  $q_n(X)$  denotes the number of points of least period  $n$  in  $X$ .

**Theorem 13** (Embedding Theorem (Krieger [70])) *Let  $X$  and  $Y$  be irreducible shifts of finite type. Then there is a proper embedding of  $X$  into  $Y$  if and only if  $h(X) < h(Y)$  and for each  $n \geq 1$ ,  $q_n(X) \leq q_n(Y)$ .*

In fact, Krieger's theorem shows that these conditions are necessary and sufficient for a proper embedding of any shift space into a mixing shift space. The analogous prob-

lems for embedding into irreducible or mixing sofic shifts are still open, though there are partial results [22].

Using the embedding theorem and other tools from symbolic dynamics, Boyle and Handelman [25] obtained a stunning application to linear algebra: namely, a complete characterization of the non-zero spectra of primitive matrices over  $\mathbf{R}$ . In fact, they obtained characterizations of non-zero spectra for primitive matrices over many other subrings of  $\mathbf{R}$ . While they did not obtain a complete characterization over  $\mathbf{Z}$ , they formulated a conjecture for  $\mathbf{Z}$  and obtained many partial results towards that conjecture, which was later proven using other tools. The result, stated below, shows that three simple necessary conditions on a set of nonzero complex numbers are actually sufficient. In order to state these conditions, we need the following notation:

Let  $\Lambda = \{\lambda_1, \dots, \lambda_k\}$  be a list of nonzero complex numbers (with multiplicity). Let  $f_\Lambda(t) = \prod_{i=1}^k (t - \lambda_i)$ , and  $\text{tr}_n(\Lambda) = \sum_{d|n} \mu(n/d) \sum_{i=1}^k \lambda_i^{n/d}$ , with  $\mu$  being the Mobius Inversion function.

- Integrality Condition:  $f_\Lambda(t)$  is a monic polynomial (with integer coefficients).
- Perron Condition: There is a positive entry in  $\Lambda$ , occurring just once, that strictly dominates in absolute value all other entries. We denote this entry by  $\lambda_\Lambda$ .
- Net Trace Condition:  $\text{tr}_n(\Lambda) \geq 0$  for all  $n \geq 1$ .

**Theorem 14** (Kim–Ormes–Roush [64]) *Let  $\Lambda$  be a list of nonzero complex numbers satisfying the Integrality, Perron, and Net Trace Conditions. Then there is a primitive integral matrix  $A$  for which  $\Lambda$  is the non-zero spectrum of  $A$ .*

These conditions are all indeed necessary for  $\Lambda$  to be the non-zero spectrum of a primitive integral matrix. The integrality condition is that  $\Lambda$  forms a complete set of algebraic conjugates; the Perron condition states that  $\Lambda$  must satisfy the conditions of the Perron–Frobenius theorem for primitive matrices; and the Net Trace condition assures that the number of periodic points of least period  $n$  would be nonnegative.

We now turn from embeddings to factors. One special case, which is somewhat related to the Road Problem above and also important for data recording applications (Sect. “Coding for Data Recording Channels”) is:

**Theorem 15** [1,81] *An SFT  $X$  factors onto the full  $n$ -shift iff  $h(X) \geq \log(n)$ .*

While this special case treats both the equal entropy case ( $h(X) = \log(n)$ ) and unequal entropy case ( $h(X) > \log(n)$ ), in general, the factor problem naturally divides into two cases: lower entropy factors and equal entropy factors. In either case, a trivial necessary condition for  $Y$  to be a factor of  $X$  is that whenever  $q_n(X) \neq 0$ , there exists a  $d/n$  such that  $q_d(Y) \neq 0$ . This condition is denoted  $P(X) \searrow P(Y)$ . Building on ideas from Krieger’s embedding theorem, this necessary condition was shown to be sufficient.

**Theorem 16** (Lower entropy factor theorem (Boyle [22])) *Let  $X$  and  $Y$  be irreducible SFT’s with  $h(X) > h(Y)$ . Then there is a factor code from  $X$  to  $Y$  if and only if  $P(X) \searrow P(Y)$ .*

As with the embedding theorem, the lower entropy factors problem for irreducible sofic shifts is still open.

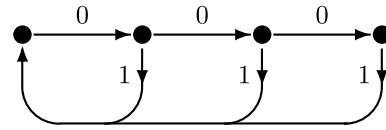
The equal entropy factors problem for SFT’s is quite different. Clearly,  $P(X) \searrow P(Y)$  is a necessary condition. A second necessary condition involves the dimension group and is simplest to state in the case of mixing edge shifts  $X_A$  and  $X_B$ .

We say that a subgroup  $\Delta$  of the dimension group  $\Delta_A$  is *pure* if whenever an integer multiple of an element  $v \in \Delta_A$  is in  $\Delta$ , then so is  $v$ ; intuitively  $\Delta$  does not have any “rational holes” in  $\Delta_A$ . The condition is that there is a pure  $\delta_A$ -invariant subgroup  $\Delta$  of  $\Delta_A$  such that  $(\Delta_B, \delta_B)$  is a quotient of  $(\Delta, \delta_A|_\Delta)$ .

In the equal entropy case, this condition and the trivial periodic point condition,  $P(X) \searrow P(Y)$ , subsume all known necessary conditions for the existence of a factor code from one mixing edge shift to another. It is not known if these two conditions are sufficient. For references on this problem, see [9,27,66].

### Coding for Data Recording Channels

In magnetic recording, within any given clock cell, a ‘1’ is represented as a change in magnetic polarity, while a ‘0’ is represented as an absence of such a change. Two successive 1’s (separated by some number,  $m \geq 0$ , of 0’s) are read as a voltage peak followed by a voltage trough (or vice versa). If the peak and trough occur too close together, *intersymbol interference* can occur: the amplitudes of the

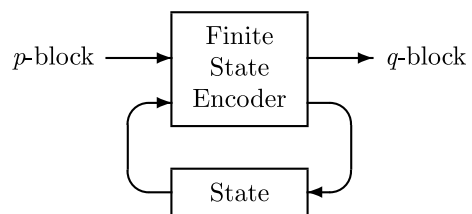


Symbolic Dynamics, Figure 7  
 $X(1, 3)$

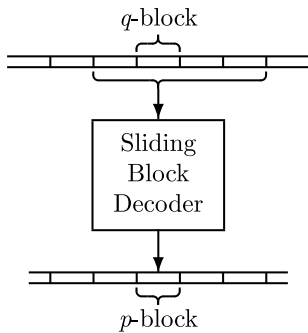
peak and trough are degraded, and the positions at which they occur are distorted. In order to control intersymbol interference, it is desirable that 1’s not be too close together, or equivalently that runs of 0’s not be too short. On the other hand, for timing control, it is desirable that runs of 0’s not be too long; this is a consequence of the fact that only 1’s are observed: the length of a run of 0’s is inferred by connection to a clock via a feedback loop, and a long run of 0’s could cause the clock to drift more than one clock cell.

This gives rise to *run-length-limited shift spaces*,  $X(d, k)$ , where runs of 0’s are constrained to be bounded below by some positive integer  $d$  and bounded above by some positive  $k \geq d$ . More precisely,  $X(d, k)$  is defined by the constraints that 1’s occur infinitely often in each direction, and there are at least  $d$  0’s, but no more than  $k$  0’s, between successive 1’s. Note that  $X(d, k)$  is an SFT with forbidden list  $\mathcal{F} = \{0^{k+1}, 10^i 1, 0 \leq i < d\}$ . Figure 7 depicts a labeled graph presentation of  $X(1, 3)$ . The SFT’s  $X(1, 3)$ ,  $X(2, 7)$ , and  $X(2, 10)$  have been used in floppy disks, hard disks, and the compact audio disk, respectively.

Now in order to record completely arbitrary information, we need to build an encoder which encodes arbitrary binary sequences into sequences that satisfy a given constraint (such as  $X(d, k)$ ). The encoder is a finite-state machine, as depicted in Fig. 8. It maps arbitrary binary data sequences, grouped into blocks of length  $p$  (called  $p$ -blocks), into constrained sequences, grouped into blocks of length  $q$  (called  $q$ -blocks). The encoded  $q$ -block is a function of the  $p$ -block as well as an internal state. When concatenated together, the sequence of encoded  $q$ -blocks must satisfy the given constraint. Also, the encoded sequences should be decodable, meaning that given the initial encoder state, a string of  $p$ -blocks can be



Symbolic Dynamics, Figure 8  
Encoder



Symbolic Dynamics, Figure 9  
Sliding block decoder

recovered from its encoded string of  $q$ -blocks, possibly allowing a fixed delay in time.

If the constrained sequences satisfy the constraints of a sofic shift  $X$ , we say that such a code is a *rate  $p:q$  finite-state code into  $X$* . In terms of symbolic dynamics, such a code consists of an edge shift  $X_G$ , a right resolving factor code  $\phi_1$  from  $X_G$  onto the full  $2^p$ -shift and a right closing sliding block code  $\phi_2$  into  $X^q$  (the right closing condition expresses the decodability condition).

In most applications, it is important that a stronger decoding condition be imposed. Namely, a *sliding block decodable rate  $p:q$  finite-state code into  $X$*  consists of a finite-state code given by  $(X_G, \phi_1, \phi_2)$  and a sliding block code  $\psi: X^q \rightarrow X_{2^p}$  such that  $\psi \circ \phi_2 = \phi_1$ . This means that the decoded  $p$ -block depends only upon the local context of the received  $q$ -block – that is, decoding is accomplished by applying a time-invariant function to a window consisting of a bounded amount of memory and/or anticipation, but otherwise is state-independent (see Fig. 9, which depicts the situation where the memory is 1 and the anticipation is 2). The point is that whenever the window of the decoder passes beyond a raw channel error, that error cannot possibly affect future decoding; thus, sliding block decoders control error propagation.

Symbolic dynamics has played an important role in providing a framework for constructing such codes as well as for establishing bounds on various figures of merit for such codes. Specifically, by modifying the constructions used in the proofs of Theorems 10 and 11, in the early 1980's, Adler, Coppersmith and Hassner (ACH) established the following results:

**Theorem 17** (*Finite-state coding theorem* [1]) *Let  $X$  be a sofic shift and  $p, q$  be positive integers. Then there is a rate  $p:q$  finite-state code into  $X$  if and only if  $p/q \leq h(X)$ .*

**Theorem 18** (*Sliding block decoding theorem* [1]) *Let  $X$  be an SFT and  $p, q$  be positive integers. Then there is a rate  $p:q$*

*sliding-block decodable finite-state code into  $X$  if and only if  $p/q \leq h(X)$ .*

We have described all of this in the context of the binary alphabet for data sequences. In fact, it works just as well for any finite alphabet, and the method used to prove Theorem 18 solved, at the same time, a special case of the factor problem for SFT's: namely, Theorem 15 above. Sometimes, Theorems 17 and 18 are stated only for rate 1:1 codes but in the context of arbitrary finite alphabets (e. g. see Chap. 5 in [78]), this easily extends to the general  $p:q$  case by passing to powers.

The ACH paper was the beginning of a rigorous theory of constrained coding. Theorem 18 has been extended to a large class of sofic shifts, and bounds have been established on such figures of merit as number of encoder states as well as the size of the decoding window of the sliding block decoder. While the state-splitting algorithm does construct codes with “relatively small” decoding windows, it is not yet understood how to construct codes with the smallest such windows. This is of substantial engineering interest since the smaller the decoding window, the smaller the error propagation. There is now a substantial literature on the construction of these types of codes, including the state-splitting algorithm as well as many other methods of encoder/decoder design and a wealth of examples that go well beyond the run length limited constraints. See for example the expositions [11,54], and [83] as well as the papers [7,8,10,12,31,41,42,43,53,56].

### Connections with Information Theory and Ergodic Theory

The concept of entropy was developed by Shannon [119] in information theory in the 1940's and was adapted to ergodic theory in the 1950's and to dynamical systems, and in particular symbolic dynamics, in the 1960's. Shannon focused on entropy for random variables and finite sequences of random variables, but the concept naturally extends to stationary stochastic processes, in particular stationary Markov chains. Roughly speaking, for a stationary process  $\mu$ , the entropy  $h(\mu)$  is the asymptotic growth rate of the number of allowed sequences, weighted by the joint stationary probabilities of  $\mu$  (for background on entropy and information theory see Cover and Thomas [35]).

A Markov chain on a graph  $G$  is defined by assigning transition probabilities to the edges. Let  $P$  denote the stochastic matrix indexed by states of  $G$ , with  $P_{IJ}$  equal to the sum of the transition probabilities of all edges from  $I$  to  $J$ . This, together with a stationary vector for  $P$ , completely defines the (joint distributions of the) Markov chain.

So, a graph itself can be viewed as specifying only which transitions are possible. For that reason, one could view  $G$  as an “intrinsic Markov chain”, and Parry [99] used this terminology when he introduced SFT’s based on graphs and matrices.

More generally, a stationary process on a shift space  $X$  is any stationary process which assigns positive probability only to allowed blocks in  $X$ .

For an irreducible graph  $G$  with strictly positive transition probabilities on all edges, by Perron–Frobenius theory,  $P$  will always have a unique stationary vector. And there is a particular such Markov chain  $\mu_G$  on  $G$  that in some sense distributes probabilities on paths as uniformly as possible. This Markov chain is the most “random” possible stationary process (not just among Markov chains) on  $X_G$  in the sense that it has maximal entropy; moreover, its entropy coincides with the topological entropy of  $X_G$ .

The Markov chain  $\mu_G$  is defined as follows: let  $\lambda$  denote the largest eigenvalue of  $A(G)$  and  $w, v$  denote corresponding left, right eigenvectors, normalized such that  $w \cdot v = 1$ ; for any path  $\gamma$  of length  $n$  from state  $I$  to state  $J$ , we define the stationary probability of  $\gamma$ :

$$\mu_G(\gamma) = \frac{w_I v_J}{\lambda^n}.$$

It is clear from the formula that this distribution is fairly uniform, since all paths with the same initial state, terminal state and length have the same stationary probability.

This defines a joint stationary distribution on allowed blocks of  $X_G$ , and it is not hard to show that it is consistent and Markov, given by assigning transition probability  $v_J/(v_I \lambda)$  to each edge from  $I$  to  $J$ . To summarize:

**Theorem 19** *Let  $G$  be an irreducible graph.*

- Any stationary process  $\mu$  on  $G$  satisfies  $h(\mu) \leq \log \lambda$ , and
- The Markov chain  $\mu_G$  is the unique stationary process on  $G$  such that  $h(\mu_G) = \log \lambda$ .

The construction of  $\mu_G$  is effectively due to Shannon [119] and uniqueness is due to Parry [99]. The unique entropy-maximizing stationary process on irreducible edge shifts naturally extends to irreducible SFT’s and irreducible sofic shifts (this process will be  $M$ -step Markov for an  $M$ -step SFT).

Many results in symbolic dynamics were originally proved using  $\mu_G$ . One example is the fact that any factor code from one irreducible SFT to another of the same entropy must be finite-to-one [32]. This result, and many others, were later proven using methods that rely only on

the basic combinatorial structure of  $G$  and  $X_G$ . Nevertheless,  $\mu_G$  provides much motivation and insight into symbolic dynamics problems, and it illustrates a connection with ergodic theory, which we now discuss.

For background on ergodic theory (such as the concepts of measure-preserving transformations, homomorphisms, isomorphisms, ergodicity, mixing, and measure-theoretic entropy), we refer the reader to [107,113,127]. Observe that a stationary process on a shift space  $X$  can be viewed as a measure-preserving transformation: the transformation is the shift mapping and the measure on “cylinder sets”, consisting of  $x \in X$  with prescribed coordinate values  $x_i = a_i, \dots, x_j = a_j$ , is defined as the probability of the word  $a_i \dots a_j$ ; the stationarity of the process translates directly into preservation of the measure. The symbolic dynamical concepts of irreducibility and mixing correspond naturally to the concepts of ergodicity and (measure-theoretic) mixing in ergodic theory.

It is well-known [107,127] that the measure-preserving transformation (MPT) defined by a stationary Markov chain  $\mu$  on an irreducible (resp., primitive) graph  $G$  is ergodic (resp., mixing) if  $\mu$  assigns strictly positive conditional probabilities to all edges of  $G$ .

Now, suppose that  $G$  and  $H$  are irreducible graphs and  $\phi: X_G \rightarrow X_H$  is a factor code. For a stationary measure  $\mu$  on  $G$ , we define a stationary measure  $\nu = \phi(\mu)$  on  $X_H$  by transporting  $\mu$  to  $X_H$ : for a measurable set  $A$  in  $X_H$ , define

$$\nu(A) = \mu(\phi^{-1}(A)).$$

Then  $\phi$  defines a measure-preserving homomorphism from the MPT defined by  $\mu$  to the MPT defined by  $\nu$ . Since measure-preserving homomorphisms between MPT’s cannot reduce measure-theoretic entropy, we have  $h(\nu) \leq h(\mu)$ .

Suppose that now  $\phi$  is actually a conjugacy. Then it defines a measure-preserving isomorphism, and so  $h(\mu) = h(\nu)$ . If  $\mu = \mu_G$ , then by uniqueness, we have  $\nu = \mu_H$ . Thus  $\phi$  defines a measure-theoretic isomorphism between the MPT defined by  $\mu_G$  and the MPT defined by  $\mu_H$ . In fact, this holds whenever  $\phi$  is merely an almost invertible factor code. This establishes the following result:

**Theorem 20** *Let  $G, H$  be irreducible graphs, and let  $\mu_G, \mu_H$  be the stationary Markov chains of maximal entropy on  $G, H$ . If  $X_G, X_H$  are almost conjugate (in particular, if they are conjugate), then the measure-preserving transformations defined by  $\mu_G$  and  $\mu_H$  are isomorphic.*

Hence conjugacies and almost conjugacies yield isomorphisms between measure-preserving transformations defined by stationary Markov chains of maximal entropy.

In fact, the isomorphisms obtained in this way have some very desirable properties compared to the run-of-the-mill isomorphism. For instance, an isomorphism  $\phi$  between stationary processes typically has an infinite window; i. e., to know  $\phi(x)_0$ , you typically need to know all of  $x$ , not just a central block  $x_{[-n,n]}$  (these are the kinds of isomorphisms that appear in general ergodic theory and in particular in Ornstein's celebrated isomorphism theory [98]). In contrast, by definition, a conjugacy always has a finite window of uniform size. It turns out that an isomorphism obtained from an almost conjugacy, as well as its inverse, has finite expected coding length in the sense that to know  $\phi(x)_0$ , you need to know only a central block  $x_{[-n(x),n(x)]}$ , where the function  $n(x)$  has finite expectation [4]. In particular, by Theorem 11, whenever  $X_G$  and  $X_H$  are mixing edge shifts with the same entropy, the measure-preserving transformations defined by  $\mu_G$  and  $\mu_H$  are isomorphic via an isomorphism with finite expected coding length.

The notions of conjugacy, finite equivalence, almost conjugacy, embedding, factor code, and so on can all be generalized to the context of stationary measures, in particular to stationary Markov chains. For instance, a conjugacy between two stationary measures is a map that is simultaneously a conjugacy of the underlying shift spaces and an isomorphism of the associated measure-preserving transformations. Many results in symbolic dynamics have been generalized to the context of stationary Markov chains. There is a substantial literature on this, in particular on finitary isomorphisms with finite expected coding time, e. g., [71,84,101,114], and [90]. The expositions [102,105] give a nice introduction to the subject of strong finitary codings between stationary Markov chains. See also the research papers [45,85,86,103,104,123,124].

## Higher Dimensional Shift Spaces

In this section we introduce higher dimensional shift spaces. For a more thorough introduction, we refer the reader to Lind [76]. For the related subject of tiling systems, see Robinson [112], Radin [110], and Mozes [94,95].

The  $d$ -dimensional full  $\mathcal{A}$ -shift is defined to be  $\mathcal{A}^{\mathbb{Z}^d}$ . Ordinarily,  $\mathcal{A}$  is a finite alphabet, and here we restrict ourselves to this case. An element  $x$  of the full shift may be regarded as a function  $x: \mathbb{Z}^d \rightarrow \mathcal{A}$ , or, more informally, as a “configuration” of alphabet choices at the sites of the integer lattice  $\mathbb{Z}^d$ .

For  $x \in \mathcal{A}^{\mathbb{Z}^d}$  and  $F \subseteq \mathbb{Z}^d$ , let  $x_F$  denote the restriction of  $x$  to  $F$ . The usual metric on the one-dimensional full shift naturally generalizes to a metric on  $\mathcal{A}^{\mathbb{Z}^d}$  given by  $\rho(x, y) = 2^{-k}$ , where  $k$  is the largest integer such that  $x_{[-k,k]^d} = y_{[-k,k]^d}$  (with the usual conventions when

$x = y$  and  $x_0 \neq y_0$ ). In analogy with one dimension, according to this definition, two points are “close” if they agree on a large cube  $[-k, k]^d$ .

We define higher dimensional shift spaces, with the following terminology. A *shape* is a finite subset  $F$  of  $\mathbb{Z}^d$ , and a *pattern*  $f$  on a shape  $F$  is a function  $f: F \rightarrow \mathcal{A}$ . We say that  $X$  is a  $d$ -dimensional shift space (or  $d$ -dimensional shift) if it can be represented by a list  $\mathcal{F}$  (finite or infinite) of “forbidden” patterns

$$X = X_{\mathcal{F}} = \{x \in \mathcal{A}^{\mathbb{Z}^d} : \sigma^n(x)_F \notin \mathcal{F} \text{ for all } n \in \mathbb{Z}^d \text{ and all shapes } F\}.$$

Just as in one dimension, we can equivalently define a shift space to be a closed (with respect to the metric  $\rho$ ) translation-invariant subset of  $\mathcal{A}^{\mathbb{Z}^d}$ . Here “translation-invariance” means that  $\sigma^n(X) = X$  for all  $n \in \mathbb{Z}^d$ , where  $\sigma^n$  is the translation in direction  $n$  defined by  $(\sigma^n(x))_m = x_{m+n}$ .

We say that a pattern  $f$  on a shape  $F$  occurs in a shift space  $X$  if there is an  $x \in X$  such that  $x_F = f$ . Hence the analogue of the language of a shift space is the set of all occurring patterns.

A  $d$ -dimensional shift of finite type  $X$  is a subset of  $\mathcal{A}^{\mathbb{Z}^d}$  defined by a finite list  $\mathcal{F}$  of forbidden patterns. Just as in one dimension, a  $d$ -dimensional shift of finite type  $X$  can also be defined by specifying allowed patterns instead of forbidden patterns, and there is no loss of generality in requiring the shapes of the patterns to be the same. Thus we can specify a finite list  $\mathcal{L}$  of patterns on a fixed shape  $F$ , and set

$$X = X_{\mathcal{L}^c} = \{x \in \mathcal{A}^{\mathbb{Z}^d} : \text{for all } n \in \mathbb{Z}^d, \sigma^n(x)_F \in \mathcal{L}\}.$$

In fact, there is no loss in generality in assuming that  $F$  is a  $d$ -dimensional cube  $F = [0, k]^d$ .

Given a finite list  $\mathcal{L}$  of patterns on a shape  $F$ , we say that a pattern  $f'$  on a shape  $F'$  is  $\mathcal{L}$ -admissible (in the shift of finite type  $X = X_{\mathcal{L}^c}$ ) if each of its sub-patterns, whose shape is a translate of  $F$ , belongs to  $\mathcal{L}$ . Of course, any pattern which occurs in  $X$  is  $\mathcal{L}$ -admissible. But an  $\mathcal{L}$ -admissible pattern need not occur in  $X$ .

The analogue of vertex shift (or 1-step shift of finite type) in higher dimensions is defined by a collection of  $d$  transition matrices  $A_1, \dots, A_d$  all indexed by the same set of symbols  $\mathcal{A} = \{1, \dots, m\}$ . We set

$$\Omega(A_1, \dots, A_d) = \{x \in \{1, \dots, m\}^{\mathbb{Z}^d} : A_i(x_n, x_{n+e_i}) = 1 \text{ for all } n, i\},$$

where  $e_i$  is as usual the  $i$ th standard basis vector and  $A_i(a, b)$  denotes the  $(a, b)$ -entry of  $A_i$ . Such a shift space



is called a *matrix subshift*. When  $d = 2$ , this amounts to a pair of transition matrices  $A_1$  and  $A_2$  with identical vertex sets. The matrix  $A_1$  controls transitions in the horizontal direction and the matrix  $A_2$  controls transitions in the vertical direction. Note that any matrix subshift is a shift of finite type and, in particular, can be specified by a list  $\mathcal{L}$  of patterns on the unit cube  $F = \{(a_1, \dots, a_n) : a_i \in \{0, 1\}\}$ ; specifically,  $\Omega(A_1, \dots, A_n) = X_{\mathcal{L}^c}$  where  $\mathcal{L}$  is the set of all patterns  $f: F \rightarrow \{1, \dots, m\}$  such that if  $n, n + e_i \in F$ , then  $A_i(f(n), f(n + e_i)) = 1$ . When we speak of admissible patterns for a matrix subshift, we mean  $\mathcal{L}$ -admissible patterns with this particular  $\mathcal{L}$ . Just as in one dimension, we can recode any shift of finite type to a matrix subshift.

Higher dimensional SFT's can behave very differently from one dimension. For example, there is a simple method to determine if a one-dimensional edge shift, and therefore a one-dimensional shift of finite type, is nonempty, and there is an algorithm to tell, for a given finite list  $\mathcal{L}$ , whether a given block occurs in  $X = X_{\mathcal{L}^c}$ . The corresponding problems in higher dimensions, called the *nonemptiness problem* and the *extension problem*, turn out to be undecidable [14,111]; see also [67]. Even for two-dimensional matrix subshifts  $X$ , these decision problems are undecidable. On the other hand, there are some special classes where these problems are decidable. This class includes any two-dimensional matrix subshift such that  $A_1$  commutes with  $A_2$  and  $A_2^\top$ . For this class, any admissible pattern on a cube must occur, and so the nonemptiness and extension problems are decidable; see [87,88].

A point  $x$  in a  $d$ -dimensional shift  $X$  is *periodic* if its orbit  $\{\sigma^n(x) : n \in \mathbb{Z}^d\}$  is finite. Observe that this reduces to the usual notion of periodic point in one dimension. Now an ordinary (one-dimensional) nonempty shift of finite type is conjugate to an edge shift  $X_G$ , where  $G$  has at least one cycle. Hence a one-dimensional shift of finite type is nonempty if and only if it has a periodic point. This turns out to be false in higher dimensions [14,111], and this fact is intimately related to the undecidability results mentioned above. While one can formulate a notion of zeta function for keeping track of numbers of periodic points, the zeta function is hard to compute, even for very special and explicit matrix subshifts, and, even in this setting, it is not a rational function [77].

In higher dimensions, the entropy of a shift is defined as the asymptotic growth rate of the number of occurring patterns in arbitrarily large cubes. In particular, for two-dimensional shifts it is defined by

$$h(X) = \lim_{n \rightarrow \infty} \frac{1}{n^2} \log |X_{[0,n-1] \times [0,n-1]}|,$$

where  $X_{[0,n-1] \times [0,n-1]}$  denotes the set of occurring pat-

terns on the square

$$[0, n-1] \times [0, n-1]$$

that occur in  $X$ . Recall from Sect. “Entropy and Periodic Points” that it is easy to compute the entropy of a (one-dimensional) shift of finite type using linear algebra. But in higher dimensions, there is no analogous formula and, in fact, other than the group shifts mentioned below, the entropies of only a very few higher dimensional shifts of finite type have been computed explicitly. Even for the two-dimensional “golden mean” matrix subshift defined by the horizontal and vertical transition matrices

$$A_1 = A_2 = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

an explicit formula for the entropy is not known. However, there are good numerical approximations to the entropy of some matrix subshifts (e.g., [30,96]). And recently the set of numbers that can occur as entropies of SFT's in higher dimensions has been characterized [51,52]; this characterization turns out to be remarkably different from the analogous characterization in one dimension [74].

One of the few 2-dimensional SFT's for which entropy has been computed is the domino tiling system (see [58], Chap. 5 in [115]), which consists of all possible tilings of the plane using the  $1 \times 2$  and  $2 \times 1$  dominoes. This can be translated into an SFT with four symbols  $L, R, T, B$ , subject to the constraints:

- $x_{i,j} = L \Rightarrow x_{i+1,j} = R$ ,
- $x_{i,j} = R \Rightarrow x_{i-1,j} = L$ ,
- $x_{i,j} = T \Rightarrow x_{i,j-1} = B$ ,
- $x_{i,j} = B \Rightarrow x_{i,j+1} = T$ .

The entropy of this SFT is given by a remarkable integral formula:

$$\frac{1}{4} \int_0^1 \int_0^1 (4 - 2 \cos(2\pi s) - 2 \cos(2\pi t)) \, ds \, dt$$

Further work along these lines can be found in [59,117].

One can formulate notions of irreducibility and mixing for higher dimensional shift spaces. It turns out that for SFT's there are several notions of mixing that all coincide in one dimension, but are vastly different in higher dimensions. For instance, a higher dimensional shift space is *strongly irreducible* if there is an integer  $R > 0$  such that for any two shapes  $F, F'$  of distance at least  $R$ , any occurring configurations on  $F$  and  $F'$  can be combined to form an occurring configuration on  $F \cup F'$  [29,129]. For one-dimensional SFT's, this is equivalent to mixing, but much stronger than mixing for two-dimensional SFT's.

Just as in one dimension, we have the notion of sliding block code for higher dimensional shifts. For finite alpha-

bets  $\mathcal{A}, \mathcal{B}$ , a cube  $F \subset \mathbb{Z}^d$  and a function  $\Phi: \mathcal{A}^F \rightarrow \mathcal{B}$ , the mapping  $\phi = \Phi_\infty: \mathcal{A}^{\mathbb{Z}^d} \rightarrow \mathcal{B}^{\mathbb{Z}^d}$  defined by

$$\Phi_\infty(x)_n = \Phi(x_{n+F})$$

is called a sliding block code. By restriction we have the notion of sliding block code from one  $d$ -dimensional shift space to another. As expected, for  $d$ -dimensional shifts  $X$  and  $Y$ , the sliding block codes  $\phi: X \rightarrow Y$  coincide exactly with the continuous translation-commuting maps from  $X$  to  $Y$ , i.e., the maps which are continuous with respect to the metric  $\rho$ , defined above, and which satisfy  $\phi \circ \sigma^n = \sigma^n \circ \phi$  for all  $n \in \mathbb{Z}^d$ . Thus it makes sense to consider the various coding problems, in particular the conjugacy, factor and embedding problems, in the higher dimensional setting, but these are very difficult.

Even the question of determining when a higher dimensional SFT of entropy at least  $\log n$  factors onto the full  $n$ -shift seems very difficult (in contrast to Theorem 15). However, there are some positive results for strongly irreducible SFT's. For instance, it is known that any strongly irreducible SFT of entropy strictly larger than  $\log n$  factors onto the full  $n$ -shift [37,55]. In fact, that result requires only a weaker assumption than strong irreducibility, but still much stronger than mixing; recent examples [28] show, among other things, that one cannot weaken that assumption to mere mixing.

There are results on other coding problems as well. For instance, a version of the Embedding theorem in one dimension (Theorem 13) holds for SFT's in two dimensions with a strong mixing property [73]; however, it is required that the shift to be embedded contains no points that are periodic in any single direction. And some other results on entropy of proper subshifts of strongly irreducible SFT's carry over from one dimension to higher dimensions [106,109]. But in many cases where versions of the result carry over, the proofs are much different from those in one dimension.

Measures of maximal entropy for two-dimensional SFT's behave very differently from the one dimensional case. For instance, even with very strong mixing properties, such as strong irreducibility, there can be more than one measure of maximal entropy [29], and the relationships among entropy-preserving, finite-to-one, and almost invertibility for factor codes discussed in Sect. "Other Coding Problems" can be very different in higher dimensions [89]. Other differences with respect to entropy can be found in [108].

There is also the natural notion of higher dimensional sofic shifts, which can be defined as those shift spaces that are factors of SFT's. Recall that every one-dimensional sofic shift is a right-resolving, and hence entropy-preserv-

ing, factor of an SFT. It is not known if there is an analogue to this fact in higher dimensions, although recently there has been some progress: every sofic shift  $Y$  is a factor of an SFT with entropy arbitrarily close to  $h(Y)$  [36].

Finally, there is a subclass of  $d$ -dimensional SFT's which is somewhat tractable, namely  $d$ -dimensional shifts with group structure in the following sense. Let  $\mathcal{A}$  be a (finite) group. Then the full  $d$ -dimensional shift over  $\mathcal{A}$  is also a group with respect to the coordinate-wise group structure. A (higher-dimensional) *group shift* is a subshift of  $\mathcal{A}^{\mathbb{Z}^d}$  which is also a subgroup. For a survey on results for this class, we refer the reader to [79].

## Future Directions

The future directions of the subject will likely be determined by progress on solutions to open problems. In the course of describing topics in this article, we have mentioned many open problems along the way. For a much more complete list on a wealth of sub-areas of symbolic dynamics, we refer the reader to the article [24]. While it is difficult to single out the most important challenges, certainly the problem of understanding multi-dimensional shift spaces, especially of finite type, is one of the most important.

## Bibliography

1. Adler RL, Coppersmith D, Hassner M (1983) Algorithms for sliding block codes – an application, of symbolic dynamics to information theory. *Trans IEEE Inf Theory* 29:5–22
2. Adler RL, Goodwyn LW, Weiss B (1977) Equivalence of topological Markov shifts. *Isr J Math* 27:48–63
3. Adler R, Konheim A, McAndrew M (1965) Topological entropy. *Trans Amer Math Soc* 114:309–319
4. Adler R, Marcus B (1979) Topological entropy and equivalence of dynamical systems. *Mem Amer Math Soc* 219. AMS, Providence
5. Adler R, Weiss B (1970) Similarity of automorphisms of the torus. *Mem Amer Math Soc* 98. AMS, Providence
6. Aho AV, Hopcroft JE, Ullman JD (1974) The design and analysis of computer algorithms. Addison-Wesley, Reading
7. Ashley J (1988) A linear bound for sliding block decoder window size. *Trans IEEE Inf Theory* 34:389–399
8. Ashley J (1996) A linear bound for sliding block decoder window size, II. *Trans IEEE Inf Theory* 42:1913–1924
9. Ashley J (1991) Resolving factor maps for shifts of finite type with equal entropy. *Ergod Theory Dynam Syst* 11:219–240
10. Béal M-P (1990) The method of poles: a coding method for constrained channels. *Trans IEEE Inf Theory* 36:763–772
11. Béal M-P (1993) *Codage symbolique*. Masson, Paris
12. Béal M-P (2003) Extensions of the method of poles for code construction. *Trans IEEE Inf Theory* 49:1516–1523
13. Bedford T (1986) Generating special Markov partitions for hyperbolic toral automorphisms using fractals. *Ergod Theory Dynam Syst* 6:325–333

14. Berger R (1966) The undecidability of the Domino problem. Mem Amer Math Soc. AMS, Providence
15. Berstel J, Perrin D (1985) Theory of codes. Academic Press, New York
16. Blanchard P, Devaney R, Keen L (2004) Complex dynamics and symbolic dynamics. In: Williams S (ed) Symbolic dynamics and its applications. Proc Symp Appl Math. AMS, Providence, pp 37–59
17. Blanchard F, Maass A, Nogueira A (2000) Topics in symbolic dynamics and applications. In: Blanchard F, Maass A, Nogueira A (eds) LMS Lecture Notes, vol 279. Cambridge University Press, Cambridge
18. Bowen R (1970) Markov partitions for Axiom A diffeomorphisms. Amer J Math 92:725–747
19. Bowen R (1973) Symbolic dynamics for hyperbolic flows. Amer J Math 95:429–460
20. Bowen R, Franks J (1977) Homology for zero-dimensional basic sets. Ann Math 106:73–92
21. Bowen R, Lanford OE (1970) Zeta functions of restrictions of the shift transformation. Proc Symp Pure Math AMS 14:43–50
22. Boyle M (1983) Lower entropy factors of sofic systems. Ergod Theory Dynam Syst 3:541–557
23. Boyle M (1993) Symbolic dynamics and matrices. In: Brualdi R et al (eds) Combinatorial and graph theoretic problems in linear algebra. IMA Vol Math Appl 50:1–38
24. Boyle M (2007) Open problems in symbolic dynamics. Contemporary Math (to appear)
25. Boyle M, Handelmann D (1991) The spectra of nonnegative matrices via symbolic dynamics. Ann Math 133:249–316
26. Boyle M, Krieger W (1986) Almost Markov and shift equivalent sofic systems. In: Alexander J (ed) Proceedings of Maryland Special Year in Dynamics 1986–87. Lecture Notes in Math, vol 1342. Springer, Berlin, pp 33–93
27. Boyle M, Marcus BH, Trow P (1987) Resolving maps and the dimension group for shifts of finite type. Mem Amer Math Soc 377. AMS, Providence
28. Boyle M, Pavlov R, Schraudner M (2008) preprint
29. Burton R, Steif J (1994) Nonuniqueness of measures of maximal entropy for subshifts of finite type. Ergod Theory Dynam Syst 14(2):213–235
30. Calkin N, Wilf H (1998) The number of independent sets in a grid graph. SIAM J Discret Math 11:54–60
31. Cidecyian R, Evangelos E, Marcus B, Modha M (2001) Maximum transition run codes for generalized partial response channels. IEEE J Sel Area Commun 19:619–634
32. Coven E, Paul M (1974) Endomorphisms of irreducible shifts of finite type. Math Syst Theory 8:167–175
33. Coven E, Paul M (1975) Sofic systems. Isr J Math 20:165–177
34. Coven E, Paul M (1977) Finite procedures for sofic systems. Monats Math 83:265–278
35. Cover T, Thomas J (1991) Elements of information theory. Wiley, New York
36. Desai A (2006) Subsystem entropy for  $\mathbb{Z}^d$  sofic shifts. Indag Math 17:353–360
37. Desai A (2008) A class of  $\mathbb{Z}^d$ -subshifts which factor onto lower entropy full shifts. Proc Amer Math Soc, to appear
38. Devaney R (1987) An introduction to chaotic dynamical systems. Addison-Wesley, Reading
39. Fischer R (1975) Sofic systems and graphs. Monats Math 80:179–186
40. Fischer R (1975) Graphs and symbolic dynamics. Colloq Math Soc János Bolyai: Top Inf Theory 16:229–243
41. Franaszek PA (1968) Sequence-state coding for digital transmission. Bell Syst Tech J 47:143–155
42. Franaszek PA (1982) Construction of bounded delay codes for discrete noiseless channels. J IBM Res Dev 26:506–514
43. Franaszek PA (1989) Coding for constrained channels: a comparison of two approaches. J IBM Res Dev 33:602–607
44. Friedman J (1990) On the road coloring problem. Proc Amer Math Soc 110:1133–1135
45. Gomez R (2003) Positive K-theory for finitary isomorphisms of Markov chains. Ergod Theory Dynam Syst 23:1485–1504
46. Hadamard J (1898) Les surfaces a courbures opposées et leurs lignes geodesiques. J Math Pure Appl 4:27–73
47. Hassellblatt B, Katok A (1995) Introduction to the modern theory of dynamical systems. Cambridge University Press, Cambridge
48. Hedlund GA (1939) The dynamics of geodesic flows. Bull Amer Math Soc 45:241–260
49. Hedlund GA (1944) Sturmian minimal sets. Amer J Math 66:605–620
50. Hedlund GA (1969) Endomorphisms and automorphisms of the shift dynamical system. Math Syst Theory 3:320–375
51. Hochman M, Meyerovitch T (2007) A characterization of the entropies of multidimensional shifts of finite type. Ann Math, to appear
52. Hochman M (2007) On the dynamics and recursive properties of multidimensional symbolic systems. Preprint
53. Hollmann HDL (1995) On the construction of bounded-delay encodable codes for constrained systems. Trans IEEE Inf Theory 41:1354–1378
54. Immink KAS (2004) Codes for mass data storage, 2nd edn. Shannon Foundation Press, Eindhoven
55. Johnson A, Madden K (2005) Factoring higher-dimensional shifts of finite type onto full shifts. Ergod Theory Dynam Syst 25:811–822
56. Karabed R, Siegel P, Soljanin E (1999) Constrained coding for binary channels with high intersymbol interference. Trans IEEE Inf Theory 45:1777–1797
57. Kari J (2001) Synchronizing finite automata on Eulerian digraphs. Springer Lect Notes Comput Sci 2136:432–438
58. Kastelyn PW (1961) The statistics of dimers on a lattice. Physica A 27:1209–1225
59. Kenyon R (2008) Lectures on dimers. <http://www.math.brown.edu/~rkenyon/papers/dimerlecturenotes.pdf>
60. Kim KH, Roush FW (1979) Some results on decidability of shift equivalence. J Comb Inf Syst Sci 4:123–146
61. Kim KH, Roush FW (1988) Decidability of shift equivalence. In: Alexander J (ed) Proceedings of Maryland special year in dynamics 1986–87. Lecture Notes in Math, vol 1342. Springer, Berlin, pp 374–424
62. Kim KH, Roush FW (1990) An algorithm for sofic shift equivalence. Ergod Theory Dynam Syst 10:381–393
63. Kim KH, Roush FW (1999) Williams conjecture is false for irreducible subshifts. Ann Math 149:545–558
64. Kim KH, Ormes N, Roush F (2000) The spectra of nonnegative integer matrices via formal power series. Amer J Math Soc 13:773–806
65. Kitchens B (1998) Symbolic dynamics: One-sided, two-sided and countable state Markov chains. Springer, Berlin
66. Kitchens B, Marcus B, Trow P (1991) Eventual factor maps

- and compositions of closing maps. *Ergod Theory Dynam Syst* 11:85–113
67. Kitchens B, Schmidt K (1988) Periodic points, decidability and Markov subgroups, dynamical systems. In: Alexander JC (ed) *Proceedings of the special year*. Springer Lect Notes Math 1342:440–454
  68. Krieger W (1980) On a dimension for a class of homeomorphism groups. *Math Ann* 252:87–95
  69. Krieger W (1980) On dimension functions and topological Markov chains. *Invent Math* 56:239–250
  70. Krieger W (1982) On the subsystems of topological Markov chains. *Ergod Theory Dynam Syst* 2:195–202
  71. Krieger W (1983) On the finitary isomorphisms of Markov shifts that have finite expected coding time. *Wahrscheinlichkeitstheorie Z* 65:323–328
  72. Krieger W (1984) On sofic systems I. *Isr J Math* 48:305–330
  73. Lightwood S (2003/04) Morphisms from non-periodic  $\mathbb{Z}^2$  subshifts I and II. *Ergod Theory Dynam Syst* 23:587–609, 24:1227–1260
  74. Lind D (1984) The entropies of topological Markov shifts and a related class of algebraic integers. *Ergod Theory Dynam Syst* 4:283–300
  75. Lind D (1989) Perturbations of shifts of finite type. *SIAM J Discret Math* 2:350–365
  76. Lind D (2004) Multi-dimensional symbolic dynamics. In: Williams S (ed) *Symbolic dynamics and its applications*. Proc Symp Appl Math 60:81–120
  77. Lind D (1996) A zeta function for  $\mathbb{Z}^d$ -actions. In: Pollicott M, Schmidt K (eds) *Proceedings of Warwick Symposium on  $\mathbb{Z}^d$ -actions*. LMS Lecture Notes, vol 228. Cambridge University Press, Cambridge, pp 433–450
  78. Lind D, Marcus B (1995) *An introduction to symbolic dynamics and coding*. Cambridge University Press, Cambridge
  79. Lind D, Schmidt K (2002) Symbolic and algebraic dynamical systems. In: Hasselblatt B, Katok A (eds) *Handbook of Dynamical Systems*. Elsevier, Amsterdam, pp 765–812
  80. Manning A (1971) Axiom A diffeomorphisms have rational zeta functions. *Bull Lond Math Soc* 3:215–220
  81. Marcus B (1979) Factors and extensions of full shifts. *Monats Math* 88:239–247
  82. Marcus BH, Roth RM (1991) Bounds on the number of states in encoder graphs for input-constrained channels. *Trans IEEE Inf Theory* 37:742–758
  83. Marcus BH, Roth RM, Siegel PH (1998) Constrained systems and coding for recording chapter. In: Brualdi R, Huffman C, Pless V (eds) *Handbook on coding theory*. Elsevier, New York; updated version at <http://www.math.ubc.ca/~marcus/Handbook/>
  84. Marcus B, Tuncel S (1990) Entropy at a weight-per-symbol and embeddings of Markov chains. *Invent Math* 102:235–266
  85. Marcus B, Tuncel S (1991) The weight-per-symbol polytope and scaffolds of invariants associated with Markov chains. *Ergod Theory Dynam Syst* 11:129–180
  86. Marcus B, Tuncel S (1993) Matrices of polynomials, positivity, and finite equivalence of Markov chains. *J Amer Math Soc* 6:131–147
  87. Markley N, Paul M (1981) Matrix subshifts for  $\mathbb{Z}^1$  symbolic dynamics. *Proc Lond Math Soc* 43:251–272
  88. Markley N, Paul M (1981) Maximal measures and entropy for  $\mathbb{Z}^1$  subshifts of finite type. In: Devaney R, Nitecki Z (eds) *Classical mechanics and dynamical systems*. Dekker Notes 70:135–157
  89. Meester R, Steif J (2001) Higher-dimensional subshifts of finite type, factor maps and measures of maximal entropy. *Pac Math J* 200:497–510
  90. Mouat R, Tuncel S (2002) Constructing finitary isomorphisms with finite expected coding time. *Isr J Math* 132:359–372
  91. Morse M (1921) Recurrent geodesics on a surface of negative curvature. *Trans Amer Math Soc* 22:84–100
  92. Morse M, Hedlund GA (1938) Symbolic dynamics. *Amer J Math* 60:815–866
  93. Morse M, Hedlund GA (1940) Symbolic dynamics II, Sturmian trajectories. *Amer J Math* 62:1–42
  94. Mozes S (1989) Tilings, substitutions and the dynamical systems generated by them. *J Anal Math* 53:139–186
  95. Mozes S (1992) A zero entropy, mixing of all orders tiling system. In: Walters P (ed) *Symbolic dynamics and its applications*. Contemp Math 135:319–326
  96. Nagy Z, Zeger K (2000) Capacity bounds for the three-dimensional (0, 1) run length limited channel. *Trans IEEE Inf Theory* 46:1030–1033
  97. Nasu M (1986) Topological conjugacy for sofic systems. *Ergod Theory Dynam Syst* 6:265–280
  98. Ornstein D (1970) Bernoulli shifts with the same entropy are isomorphic. *Adv Math* 4:337–352
  99. Parry W (1964) Intrinsic Markov chains. *Trans Amer Math Soc* 112:55–66
  100. Parry W (1977) A finitary classification of topological Markov chains and sofic systems. *Bull Lond Math Soc* 9:86–92
  101. Parry W (1979) Finitary isomorphisms with finite expected code-lengths. *Bull Lond Math Soc* 11:170–176
  102. Parry W (1991) Notes on coding problems for finite state processes. *Bull Lond Math Soc* 23:1–33
  103. Parry W, Schmidt K (1984) Natural coefficients and invariants for Markov shifts. *Invent Math* 76:15–32
  104. Parry W, Tuncel S (1981) On the classification of Markov chains by finite equivalence. *Ergod Theory Dynam Syst* 1:303–335
  105. Parry W, Tuncel S (1982) Classification problems in ergodic theory. In: LMS Lecture Notes, vol 67. Cambridge University Press, Cambridge
  106. Pavlov R (2007) Perturbations of multi-dimensional shifts of finite type. Preprint
  107. Petersen K (1989) *Ergodic theory*. Cambridge University Press, Cambridge
  108. Quas A, Sahin A (2003) Entropy gaps and locally maximal entropy in  $\mathbb{Z}^d$ -subshifts. *Ergod Theory Dynam Syst* 23:1227–1245
  109. Quas A, Trow P (2000) Subshifts of multidimensional shifts of finite type. *Ergod Theory Dynam Syst* 20:859–874
  110. Radin C (1996) Miles of tiles. In: Pollicott M, Schmidt K (eds) *Ergodic Theory of  $\mathbb{Z}^d$ -actions*. LMS Lecture Notes, vol 228. Cambridge University Press, Cambridge, pp 237–258
  111. Robinson RM (1971) Undecidability and nonperiodicity for tilings of the plane. *Invent Math* 12:177–209
  112. Robinson EA (2004) Symbolic dynamics and tilings of  $\mathbb{R}^d$ . In: Williams S (ed) *Symbolic dynamics and its applications*. Proc Symp Appl Math 60:81–120
  113. Rudolph D (1990) *Fundamentals of measurable dynamics*. Oxford University Press, Oxford



114. Schmidt K (1984) Invariants for finitary isomorphisms with finite expected code lengths. *Invent Math* 76:33–40
115. Schmidt K (1990) Algebraic ideas in ergodic theory. *AMS-CBMS Reg Conf* 76
116. Schmidt K (1995) *Dynamical systems of algebraic origin*. Birkhauser, Basel
117. Schwartz M, Bruck S (2008) Constrained codeds as networks of relations. *IEEE Trans Inf Theory* 54:2179–2195
118. Seneta E (1980) Non-negative matrices and Markov chains, 2nd edn. Springer, Berlin
119. Shannon C (1948) A mathematical theory of communication. *Bell Syst Tech J* 27:379–423,623–656
120. Sinai YG (1968) Markov partitions and C-diffeomorphisms. *Funct Anal Appl* 2:64–89
121. Smale S (1967) Differentiable dynamical systems. *Bull Amer Math Soc* 73:747–817
122. Trachtman A (2007) The road coloring problem. *Israel J Math*, to appear
123. Tuncel S (1981) Conditional pressure and coding. *Isr J Math* 39:101–112
124. Tuncel S (1983) A dimension, dimension modules and Markov chains. *Proc Lond Math Soc* 46:100–116
125. Wagoner J (1992) Classification of subshifts of finite type revisited. In: Walters P (ed) *Symbolic dynamics and its applications*. *Contemp Math* 135:423–444
126. Wagoner J (2004) Strong shift equivalence theory. In: Walters P (ed) *Symbolic dynamics and its applications*. *Proc Symp Appl Math* 60:121–154
127. Walters P (1982) *An introduction to ergodic theory*. Springer Grad Text Math 79. Springer, Berlin
128. Walters P (1992) *Symbolic dynamics and its applications*. In: Walter P (ed) *Contemp Math* 135. AMS, Providence
129. Ward T (1994) Automorphisms of  $\mathbb{Z}^d$ -subshifts of finite type. *Indag Math* 5:495–504
130. Weiss B (1973) Subshifts of finite type and sofic systems. *Monats Math* 77:462–474
131. Williams RF (1973/74) Classification of subshifts of finite type. *Ann Math* 98:120–153; Erratum: *Ann Math* 99:380–381
132. Williams S (2004) Introduction to symbolic dynamics. In: Williams S (ed) *Symbolic dynamics and its applications*. *Proc Symp Appl Math* 60:1–12
133. Williams S (2004) *Symbolic dynamics and its applications*. In: Williams S (ed) *Proc Symp Appl Math* 60. AMS, Providence

## Synchronization Phenomena on Networks

GUANRONG CHEN<sup>1</sup>, MING ZHAO<sup>2</sup>, TAO ZHOU<sup>2</sup>,  
BING-HONG WANG<sup>2,3</sup>

<sup>1</sup> Department of Electronic Engineering,  
City University of Hong Kong, Hong Kong, China

<sup>2</sup> Department of Modern Physics and Nonlinear Science  
Center, University of Science and Technology of China,  
Hefei Anhui, China

<sup>3</sup> Institute of Complex Adaptive Systems,  
Shanghai Academy of System Science, Shanghai, China

## Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Basic Concepts of Network Synchronization](#)

[Synchronizability Versus Structure](#)

[Enhancing Network Synchronizability](#)

[Future Research Outlook](#)

[Acknowledgments](#)

[Bibliography](#)

## Glossary

**Synchronization** A problem in time-keeping, requiring the coordination of events to operate a system or a task in unison.

**Distance** A measure between two nodes, defined as the number of edges connecting them through the shortest paths.

**Average distance** The mean distance, averaged over all pairs of nodes on the network.

**Clustering coefficient** The probability that two randomly-selected neighboring nodes of a node are directly connected each other.

**Node-degree** The number of edges incident from a node.

**Random-graph network** A type of graph obtained by starting with a set of nodes and then adding edges between them at random.

**Small-world network** A type of graph in which most nodes are not neighbors of each other, but most nodes can be reached from any other node by a small number of connection steps; thus, a small-world network is highly clustered like a regular graph, and yet with a small average distance, just like a random graph.

**Scale-free network** A type of graph in which a small number of nodes have a large number of connections while a large number of nodes have a small number of connections, whose node-degree distribution typically follows a power-law form, with both structure and dynamics being independent of the network size.

**Node-betweenness** A measure of the extent to which a given node is occupied by the amount of information passing through it via shortest paths between other nodes, namely, the portion of shortest paths between all pairs of nodes which have data traffic going through this particular node in the network.

## Definition of the Subject

The subject under consideration is synchronization on complex networks, with respect to the phenomena and



particularly the ability of achieving synchrony of a network of dynamical systems. The subject of synchronization is quite old, but it is a significant one continuously calling for serious and systematic investigation. Ever since the careful study of two synchronous pendulum clocks by the great Dutch scientist Christian Huygens in 1665, the subject has evolved to be an independent and indispensable field of scientific research. The current study of complex networks, on the other hand, is pervading all kinds of sciences, ranging from physical to biological, even to social sciences. Its impact on modern engineering and technology is prominent and will be far-reaching. Typical complex dynamical networks include the Internet, the World Wide Web, various wireless communication networks, metabolic networks, biological neural networks, social relationship networks, financial and economic networks, and so on. As it has turned out today, the study of synchronization phenomena and synchronous behaviors of dynamical systems such as oscillators on complex networks has become overwhelming. This article offers an overview of the state-of-the-art advances and developments of the subject of synchronization on various complex networks, with emphasis on network synchronizability and performance.

## Introduction

Many biological, social and technological systems can be properly described by complex networks with nodes representing individuals or organizations and edges characterizing the interactions among them [1,2,3,4,5]. One of the goals in the current studies on complex networks is to understand and explain how the topological properties of a network affect the behaviors of dynamical systems built upon the network. Typical examples include understanding how the topology of the Internet affects the spread of the computer viruses [6,7,8,9,10], how the structure of a power grid affects the cascading failures over time [11,12,13,14,15], how the connecting patterns of an intercommunication network affect its data traffic and dynamics [16,17,18,19,20], and so on.

Synchronous behaviors have been observed in various complex networks in nature and human society [23,24,25,26], and they have been studied for hundreds of years since the systematic investigation of pendulum synchrony by the great Dutch scientist Christian Huygens in 1665 [27].

To understand how network structure affects the synchronizability of a network not only has broad theoretical interest [28], but also has important practical value [29]. One typical case in point is the synchronicity of sen-

sors in biological neural networks, where neurons communicate with each other through synaptic junctions for which a mechanism called asynchronous release is important [30]. There are many careful studies about collective synchronization in the earlier literature, with a basic assumption that dynamical systems of coupled oscillators evolve either on regular networks [31,33,34] or on random networks [35,36]. However, the structures of most real-world networks are neither completely regular nor completely random, but rather, somewhere in between. Thus, it becomes important and even necessary to consider how network structure affects the synchronization process and the synchronizability of the dynamical systems on such networks. Recently, it has been found that networks with small-world effects and scale-free properties are quite different from, and oftentimes achieve synchronization more easily than, regular networks such as lattices [37,38,39,40,41,42,43,44].

The study of synchronization on complex networks has gone through several stages in the past decade, encompassing several important aspects of the subject: various synchronization phenomena on complex networks and their stability analysis, the relationships between structural ingredients and a network's synchronizability, the enhancement or reduction of network synchronizability, etc. The first two are quite well understood today while the last one will be further addressed in this article. First, some basic concepts about synchronization of networked dynamical systems and the associated stability analysis are introduced. Second, some intrinsic relations between network structure and synchronizability are discussed. Third, three types of methods, namely, regulating coupling patterns, modifying network structures, and designing output functions, are introduced for enhancing network synchronizability. Finally, some open questions are posed which are deemed significant for further studies of the important subject of complex network synchronization.

To proceed, some notations are introduced [5], among which three are most significant with respect to network synchronization: average distance  $L$ , clustering coefficient  $C$ , node-degree  $k_i$  of node  $i$ , and the corresponding probability density function of degree distribution  $p(k)$ .

In the past few years, by taking advantage of both high-speed computing power and the huge amount of real data available on the web, scientists were able to search and find some common statistical characteristics shared by many real-world networks. It is found that most real networks have a very small average distance, scaled approximately as  $L \sim \ln N$ , where  $N$  is the size of the network (i. e., the total number of its nodes), while their clustering coefficient is rather large, as compared with *random-graph networks*

(Erdős and Rényi [28]). A network having both of these two characteristics is referred to as a *small-world network*, described by Watts and Strogatz [21]. Moreover, the degree distributions of many real networks obey a power-law form  $p(k) \sim k^{-\gamma}$ , where  $p(k)$  is the probability density function for the corresponding degree distribution, and  $\gamma$  is the power-law exponent (typically  $2 < \gamma < 3$ ) [1,2,3]. The power-law distribution falls off much more gradually than an exponential one, allowing for a few nodes with very large degrees to exist. Networks with power-law degree distributions usually belong to the class of *scale-free networks*, characterized by Barabási and Albert [22].

### Basic Concepts of Network Synchronization

A general model of coupled identical oscillators on a network can be described by [40,41]

$$\dot{\mathbf{x}}_i = \mathbf{F}(\mathbf{x}_i) - \sigma \sum_{j=1}^N G_{ij} \mathbf{H}(\mathbf{x}_j), \quad i = 1, \dots, N, \quad (1)$$

where  $\dot{\mathbf{x}}_i = \mathbf{F}(\mathbf{x}_i)$  governs the dynamics of the  $i$ th oscillator, with state vector  $\mathbf{x}_i$ ;  $\mathbf{H}(\mathbf{x}_j)$  is the output function;  $\sigma$  is the coupling strength;  $G = [G_{ij}]$  is an  $N \times N$  coupling matrix determined by the given coupling pattern among the  $N$  oscillators.

In the typical situation when the oscillators are symmetrically coupled, the coupling matrix  $G$  has the same form as the graph Laplacian  $L$ , i. e.,  $G = L$ , with

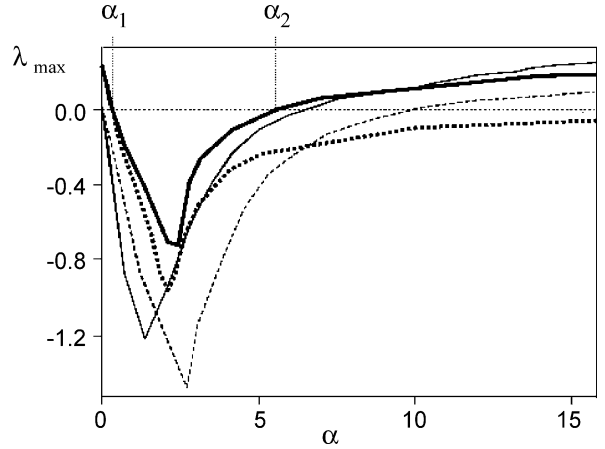
$$L_{ij} = \begin{cases} k_i & \text{for } i = j \\ -1 & \text{for } j \in \Lambda_i \\ 0 & \text{otherwise,} \end{cases} \quad (2)$$

where  $k_i$  is the degree of node  $i$  and  $\Lambda_i$  is the set of its neighboring nodes. In this setting,  $L$  is symmetrical and semi-positive definite, and all the rows of  $L$  have a zero sum, so that its smallest eigenvalue  $\lambda_1$  is always a single zero and all the other eigenvalues are strictly positive. Thus, the eigenvalues of  $L$  can be ranked as

$$0 = \lambda_1 < \lambda_2 \leq \lambda_3 \leq \dots \leq \lambda_N.$$

For network (1), the synchronization manifold is an invariant manifold:  $\mathbf{x}_1 = \mathbf{x}_2 = \dots = \mathbf{x}_N = \mathbf{s}$ , typically satisfies  $\dot{\mathbf{s}} = \mathbf{F}(\mathbf{s})$  in engineering applications.

For a dynamical system, the so-called master stability function is usually defined to be the ratio of the largest Lyapunov exponent versus a connectivity parameter of the system [42,45]. For some dynamical systems, the master stability function is negative when  $\lambda_2 > \alpha_1/\sigma$  for some



**Synchronization Phenomena on Networks, Figure 1**

Four typical master stability functions for coupled Rössler oscillators: chaotic (bold curve) and periodic (regular curve); with y-coupling (dashed curve) and x-coupling (dotted curve). The vertical ordinate shows the change of the largest Lyapunov exponent. Curves are all scaled for clearer visualization (after [42])

constant  $\alpha_1$ . In this case, the largest Lyapunov exponent is negative, and consequently the network is synchronizable; moreover, the larger the  $\lambda_2$  is, the better the network synchronizability will be [40,41].

For some other dynamical systems, the master stability function is negative only within a finite interval  $(\alpha_1, \alpha_2)$  [46], over which the largest Lyapunov exponent is negative [42,45], where  $\alpha_1$  and  $\alpha_2$  are constants. In this case, the network is synchronizable for some  $\sigma$  when the eigenratio  $R = \lambda_N/\lambda_2$  satisfies  $R < \alpha_2/\alpha_1$ ; moreover, a smaller  $R$  indicates a better network synchronizability.

The former case corresponds to networks for which the synchronized region is unbounded (the bold-dashed curve in Fig. 1), and the latter, bounded (the bold-solid curve and the two regular lines in Fig. 1) [42]. In both cases, the right-hand side of the above two inequalities depends only on the dynamics of each individual oscillator and the output function of the network, while the eigenvalue  $\lambda_2$  and eigenratio  $R$  depend only on the Laplacian  $L$ . Therefore, the problem of synchronization can be divided into two parts: choosing suitable dynamics (including the aforementioned parameters and output function) and analyzing the eigenvalues of the Laplacian. In fact, these two cases can co-exist [32,101].

The same stability analysis can also be applied to some more complicated coupling patterns [40,41,42,45,47], including the case where  $G$  is non-diagonalizable (see Fig. 2 and [48]).

Network (1) has only identical oscillators, while in the real world parameter mismatch between oscillators is very

common, so that both the amplitudes and phases of different oscillators become different. However, quite often, only the frequencies of oscillations are of concern in some applications, while the amplitudes are not important. In such cases, phase synchronization is the topic for study, for which the *Kuramoto model* [49,50,51,52,53] is a representative platform.

In the Kuramoto model, oscillators run at arbitrary frequencies and they are coupled through a periodic (e.g., sine) function of their phase differences. More precisely, the model consists of a population of  $N$  coupled phase-oscillators  $\theta_i(t)$  having natural frequencies  $\omega_i$  distributed with a given probability density  $g(\omega)$ , governed by

$$\dot{\theta}_i = \omega_i - \sigma \sum_{j=1}^N G_{ij} \sin(\theta_i - \theta_j), \quad i = 1, \dots, N. \quad (3)$$

To measure the synchronization phenomena, an order parameter  $M$  is introduced:

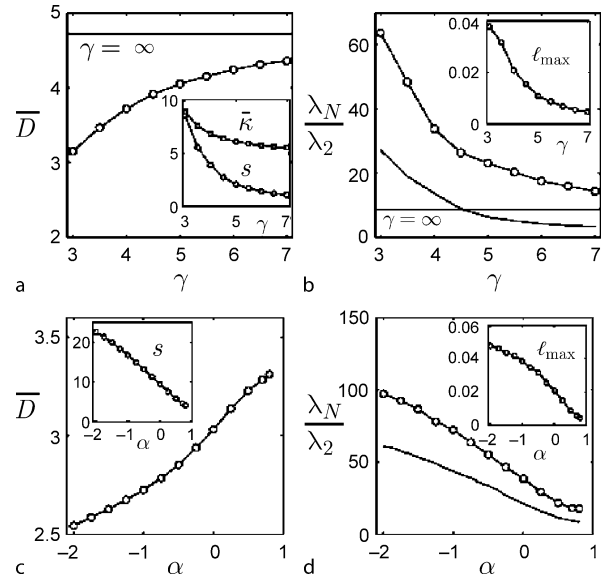
$$M \equiv \left[ \left| N^{-1} \sum_{j=1}^N e^{i\phi_j} \right| \right], \quad (4)$$

where  $\phi$  is a function of  $\theta$ , and  $\langle \cdot \rangle$  and  $[\cdot]$  denote the average over time and over different configurations, respectively.

Initially, each node is assigned a random phase. Without coupling, all the oscillators run independently and, at any time, the phases of the oscillators are distributed almost uniformly on the interval  $[0, 2\pi]$ , yielding  $M = O(1/\sqrt{N})$ . In this situation, the oscillators are generally not synchronized. With coupling, as the coupling strength gradually increases to beyond a certain threshold, interactions among oscillators become stronger and more inter-influential, which gradually dominate the individual self-oscillations. Eventually, collective synchronization of all oscillators emerges spontaneously. During this transition process, the order parameter  $M$  increases from 0 to 1.

### Synchronizability Versus Structure

Previous studies have demonstrated that both scale-free and small-world networks are much easier to synchronize than regular lattices [37,38,39,40,41,42,43,44]. At this point, a natural question arises: what makes them easier to synchronize? An intuitive answer might be their average distance, which is much shorter than that of a regular network with the same size. However, after some systematic investigations on the relation between structural ingredients and the network synchronizability, Nishikawa et al. [54] found that as the network becomes more heterogeneous, i.e., the degree distribution becomes wider,



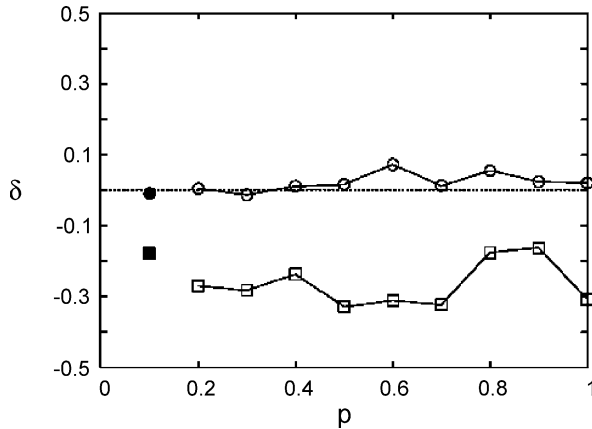
**Synchronization Phenomena on Networks, Figure 2**

**Synchronization of scale-free networks.** a, b the semi-random model; c, d the growing model with aging of nodes. The small insets are the responses of the indicated parameters with respect to the changing parameters  $\gamma$  or  $\alpha$  under the same conditions (after [54])

a network can become less synchronizable even though its average distance becomes much shorter. Figure 2 gives two examples of this phenomenon. In a semi-random model [55], with the increase of the power-law exponent  $\gamma$ , which makes the network more homogeneous, the network average distance  $\bar{D}$  becomes longer and the standard deviation of the degree distribution reduces (Fig. 2a and inset); meanwhile, the eigenratio  $\lambda_N/\lambda_2$  of its Laplacian becomes smaller (Fig. 2b), indicating improvement of the network synchronizability. In a growing model of scale-free networks with aging nodes [56], it is also observed that as the average distance increases and the degree distribution becomes more homogeneous, the network gains a better synchronizability (Fig. 2c,d).

A heuristic exploration may be given: in a network with a heterogeneous degree distribution, a few “central” oscillators, which interact with a large number of other oscillators, tend to be overloaded by the traffic passing through them. When too many independent traffic signals with different phases and frequencies are traversing through a node at the same time, they cause congestion, leading to the reduction of network synchronizability. The same also happens to overloaded edges [54].

On the other hand, based on experience with WS small-world networks, Hong et al. [57] concluded that the

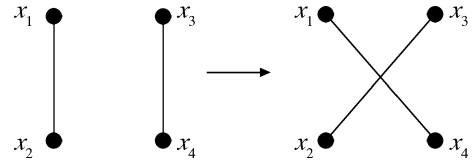


**Synchronization Phenomena on Networks, Figure 3**

Behavior of the difference  $\delta$  of the eigenratio in a WS network with rewiring probability  $p$  (after [57])

maximal node betweenness [58,59,60] is a good indicator for network synchronizability: the smaller, the better, and vice versa. To confirm their observation, they calculated the difference of the eigenratio before and after the removal of a node from a WS network [21]. Figure 3 plots the difference  $\delta \equiv (\lambda_N/\lambda_2)_{\text{after}} - (\lambda_N/\lambda_2)_{\text{before}}$ . The reduction of the ratio is brought about by the removal of the node with the maximal betweenness (empty squares in the figure). In comparison, random removal of a node makes the eigenratio almost unchanged (empty circles in the figure). This implies that the node with the maximal betweenness plays an important role in determining the synchronizability of the network. However, for scale-free networks, this “maximal betweenness indicator” may not work, as pointed out in [61] with a counterexample given in [62].

In the above studies, a network is usually modified in order to see how the synchronizability changes as the network structure varies. It is worth emphasizing that during the modification process all the topological ingredients [5] have been changed at the same time, therefore it is impossible to obtain any accurate relation between one particular ingredient and the network synchronizability. Knowing this problem, by using the edge-exchange operation [63,64], Zhao et al. [62] derived some fairly accurate relations between the synchronizability and the average distance as well as the heterogeneity of the degree distribution, on small-world and scale-free network models. Figure 4 presents a sketch of maps of their random interchanging algorithms. The algorithmic operations will change only the network average distance while keeping the degree of each node unchanged. Thus, the relations between the two concerned ingredients can be investigated



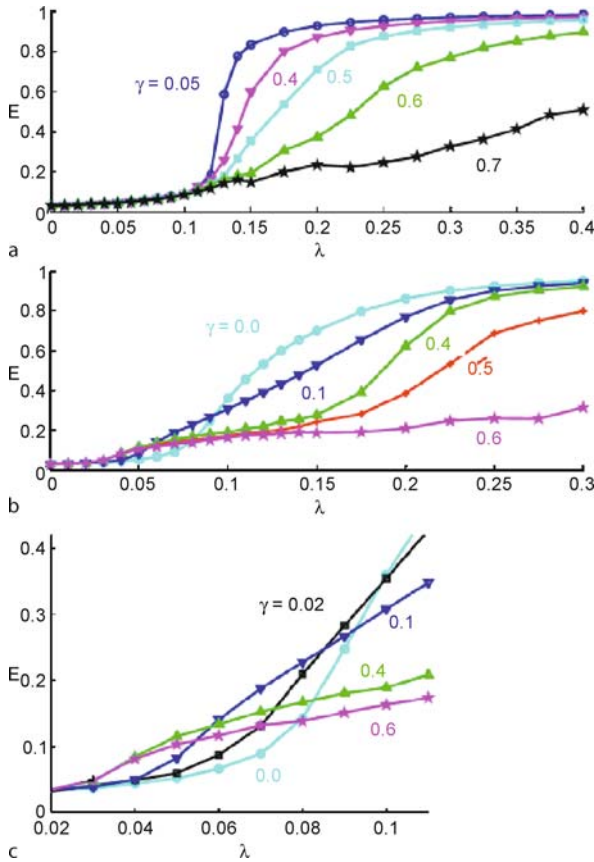
**Synchronization Phenomena on Networks, Figure 4**

Sketch of maps of the random interchanging algorithm (after [62])

separately. Extensive simulations have verified that either shortening the average distance or lowering the heterogeneity may lead to a better synchronizability, but only their combination can always ensure that the network will synchronize easily.

McGraw and Menzies [65] investigated the relations between the clustering coefficient and network synchronizability, and concluded that for both random-graph and scale-free networks, increasing the clustering coefficient hinders global synchronization if the coupling strength is strong, but it promotes the synchronization of scale-free networks when the coupling strength is weak. Figure 5 shows this phenomenon. The main reason is that the clusters around the hub-nodes promote the formation of frequency-synchronized clusters, but they will inhibit the synchronization of the network as a whole. The early hub synchronization accounts for the slightly enhanced order parameter when the coupling is weak [65,66]. This analysis is based on non-identical oscillators in the Kuramoto model. On the other hand, by means of master stability analysis, Wu et al. [67] reported a negative correlation between the clustering coefficient and synchronizability through a scale-free network model with a tunable clustering coefficient [68].

Besides the main focus on small-world effects and scale-free properties, as described by the clustering coefficient, average distance and degree distribution, some further studies on the effects of other topological ingredients on network synchronization have also been reported, particularly the degree-degree correlation. A network is said to show *assortative* (or *disassortative*) *mixing*, if the nodes having many connections tend to connect to other nodes with many (or few) connections. The extent of this degree-degree correlation can be measured by the Pearson coefficient [69]: its positive (or negative) value indicates assortative (or disassortative) mixing. Di Bernardo et al. found that disassortative networks generally have a better synchronizability than the assortative ones [70,71]. However, later works [72,73] show that the degree distribution, coupling pattern, and degree-degree correlation among the nodes compete with each other in an intrinsic manner,

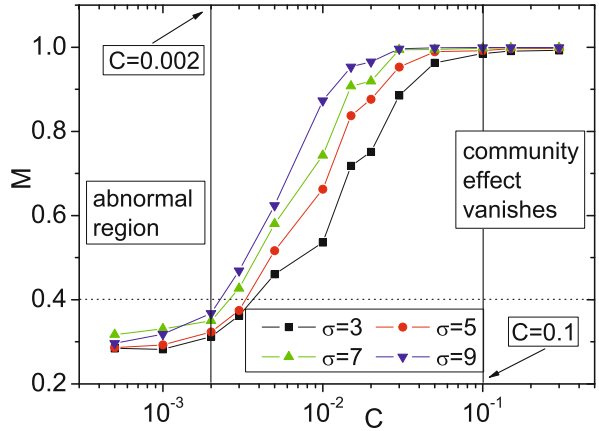


**Synchronization Phenomena on Networks, Figure 5**

Order parameter  $M$  vs coupling strength  $\lambda$ , for different values of the clustering coefficient  $\gamma$ . **a** Poisson degree distribution. **b, c** Power-law degree distribution. **c** A close-up of the transition region, showing that increase of the clustering coefficient leads to an advanced (lower- $\lambda$ ) transition (after [65])

thereby together determining the network synchronizability. That is, for one coupling pattern, disassortative mixing may predict better synchronizability, while for another coupling pattern, the result can be the opposite.

As we gain more knowledge of various network structures, more attention is paid to the effects of local structures of complex networks on their global behaviors and dynamics. Huang et al. [74] found that in complex networks with prominent clusters, the synchronizability is determined by the interplay between intercluster and intracluster edges: a network is mostly synchronizable when the numbers of the two types of edges are approximately equal. If not equal, for example as the number of intracluster edges increases, an abnormal synchronization phenomenon appears: although the network average distance becomes smaller, the network synchrony is weakened or even destroyed.



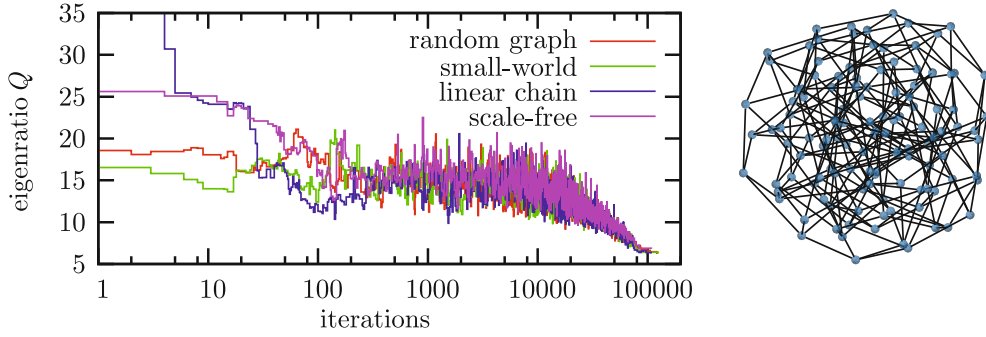
**Synchronization Phenomena on Networks, Figure 6**

Order parameter  $M$  vs. community strength  $C$  for different values of the coupling strength  $\sigma$  (after [76])

Furthermore, the synchronization phenomenon of a complex network with a community structure has also been discussed. Qualitatively, a *community* is defined as a subset of nodes within a network with the property that the connections among the nodes therein are denser than those within the other parts of the network [75]. Zhou et al. studied phase synchronization in a network with a community structure [76]. Defining the edges connecting two nodes in one community as *internal edges*, and those connecting nodes between two communities as *external edges*, the ratio of the number of external edges to the number of internal edges can be used to characterize the strength of the community structure, denoted by  $C$ . Clearly, a smaller  $C$  corresponds to sparser external edges thus a more prominent community structure. Figure 6 shows the relationship between the order parameter  $M$  and the community strength  $C$  for different coupling strengths  $\sigma$ . It is found from Fig. 6 that a strong community structure will hinder global synchronization no matter what the coupling strength is, but this effect will vanish when the fraction of external connections exceeds 0.1.

Using a modified simulated annealing algorithm, Donetti et al. [77] generated an entangled network with optimal synchronizability. These kinds of networks are shown to have an extremely homogeneous structure: distributions of node degrees, distances, betweenness, and loops are all very uniform. Also, these networks are characterized by short average distances and large loops, with no well-defined community structures. In the approach of [77], rewiring is applied, i. e., at each time step, the number of rewiring trials is randomly extracted from an exponential distribution. Except for rewiring which reduces or increases the eigenratio, and except for operations that





**Synchronization Phenomena on Networks, Figure 7**

Eigenratio  $Q$  as a function of the number of algorithmic iterations. Starting from different initial configurations, all the networks are converted via iterations to *entangled networks* (after [77])

disconnect the network, for different initial configurations the optimization process will always lead to the same optimal result. Figure 7 shows the changes of the eigenratio in the optimizing process and the resultant network configuration.

### Enhancing Network Synchronizability

With a clearer understanding of the relations between the network structure and synchronizability, a natural question about how to enhance the network synchronizability is in order. Some effective synchronizability-enhancement methods are introduced in this section.

#### Coupling Pattern Regulation

In general, scale-free networks are much harder to synchronize than random networks with the same size and the same average degree. One reason is that in scale-free networks, there are some “central” oscillators that interact with a large number of other nodes [54]. Thus, when too many independent signals with different phases and frequencies are traversing through a “central” oscillator at the same time they may have conflicts, thereby causing traffic congestion. Hence, generally speaking, the more heterogeneous the degree distribution, the more difficult for the network to synchronize. It is also known that in scale-free networks, when the oscillators are coupled symmetrically, oscillators with larger degrees usually approach the final synchronized state first, and then the others with smaller degrees synchronize to them gradually [65]. Therefore, when the oscillators are coupled asymmetrically, if the coupling strength from the “central” oscillators to the other nodes are stronger than the reverse, the network will synchronize much easier and faster.

Based on this idea, Motter, Zhou and Kurths [78,79,80] proposed a new coupling pattern, which we will call the

MZK pattern, which can sharply improve network synchronizability. After that, quite a few methods for regulating coupling patterns are brought forward to improve network synchronizability, some static and some dynamic.

#### Static Coupling Patterns

In static coupling patterns, the elements of the coupling matrix are formulated based on the MZK pattern, as

$$G_{ij} = L_{ij}/k_i^\beta, \quad (5)$$

where  $\beta$  is a tunable parameter. The coupling is weighted when  $\beta \neq 0$ , and unweighted when  $\beta = 0$ . In spite of the asymmetry of this coupling matrix  $G$ , it can be proved that all the eigenvalues of  $G$  are nonnegative reals with only one eigenvalue being zero if the network is connected. Rewrite Eq. (5) as

$$G = D^{-\beta} L, \quad (6)$$

where  $D = \text{diag}(k_1, \dots, k_N)$  is a diagonal matrix and  $L$  is the Laplacian. From the identity

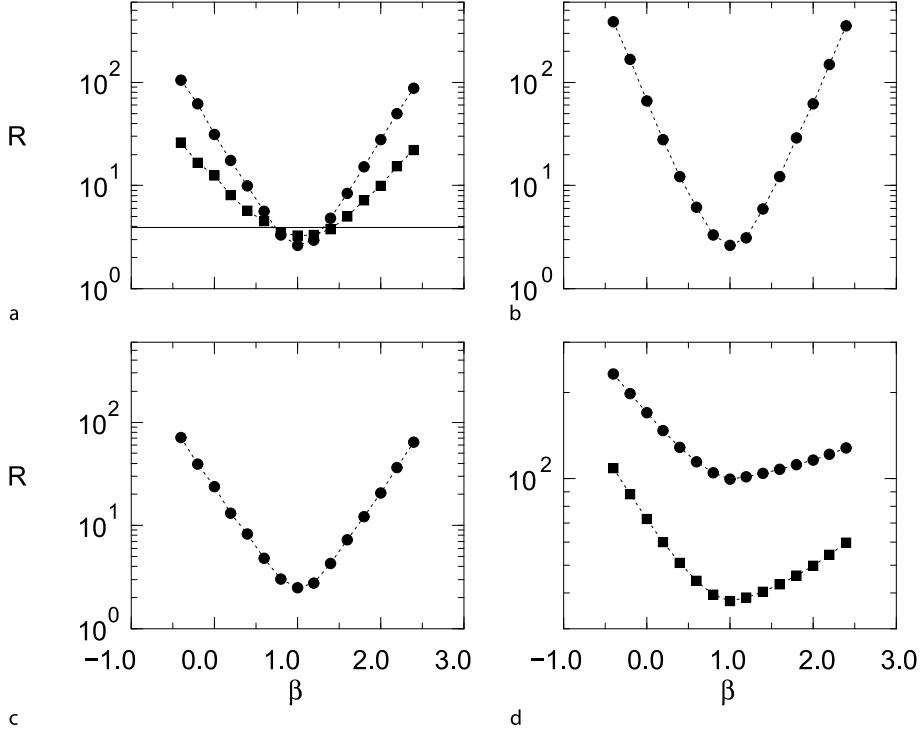
$$\det(D^{-\beta} L - \lambda I) = \det(D^{-\beta/2} L D^{-\beta/2} - \lambda I), \quad (7)$$

where  $I$  is the  $N \times N$  identity matrix, one can prove that the spectrum of  $G$  is the same as that of the following symmetric matrix:

$$H = D^{-\beta/2} L D^{-\beta/2}. \quad (8)$$

Similarly to the case of matrix  $G$ , if the network is connected then all eigenvalues of  $H$  other than the single  $\lambda_1 = 0$ , are positive. With  $\beta = 1$ , the matrix  $H$  is a normalized Laplacian. Thus, if the network is connected and  $N \geq 2$ , then

$$0 < \lambda_2 \leq N/(N-1), \quad 2 \leq \lambda_N \leq N/(N-1). \quad (9)$$



Synchronization Phenomena on Networks, Figure 8

Eigenratio  $R$  as a function of  $\beta$  for four kinds of complex networks specified in [78]. For each model, the synchronizability peaks at  $\beta = 1.0$  (after [78])

Figure 8 shows the changes of the eigenratio  $R$  with the parameter  $\beta$  in four kinds of complex networks specified in [78]. It can be seen that the eigenratio  $R$  has a well-defined minimum at  $\beta = 1$  in all cases. Mathematically, this means that the best results are obtained when the matrix  $D$  has a square-root. It is also clear that the more heterogeneous the network is, the more prominent the minimum of the eigenratio  $R$  becomes.

By explicitly relating the asymmetry in the connections to an age order among different nodes, Hwang et al. [81] found that age-ordered networks provide a better propensity for synchronization. The main reason is that an older node becomes weaker, therefore more easily influenced by other nodes. In this coupling pattern, the off-diagonal entries of the zero-row-sum coupling matrix  $G$  are

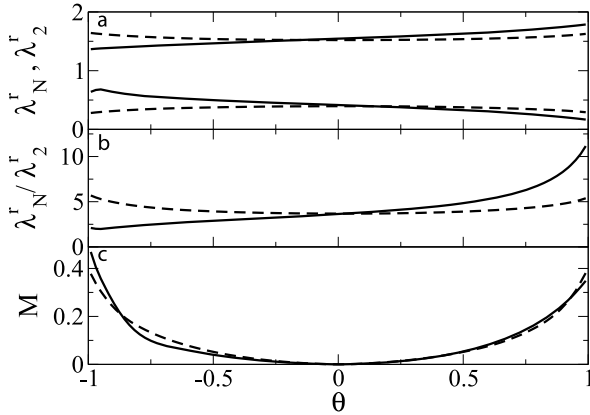
$$G_{ij} = -a_{ij} \frac{\Theta_{ij}}{\sum_{j \in \Lambda_i} \Theta_{ij}}, \quad (10)$$

where  $a_{ij}$  are the elements of the adjacency matrix  $A$  ( $a_{ij} = 1$  if nodes  $i$  and  $j$  are connected, and  $a_{ij} = 0$  otherwise), and  $\Theta_{ij} = (1 - \theta)/2$  (or  $\Theta_{ij} = (1 + \theta)/2$  for  $i > j$  (or  $i < j$ )). The parameter  $\theta \in (-1, 1)$  governs the coupling asymmetry in the network: the limit  $\theta \rightarrow -1$  (or  $\theta \rightarrow 1$ ) gives a unidirectional coupling, where the old (or

young) nodes drive the young (or old) ones. When  $\theta = 0$ , the coupling pattern degenerates to the MZK pattern at  $\beta = 1$ .

For a generic  $\theta$ , the spectrum of the coupling matrix  $G$  is in the complex plane and the complex eigenvalues appear in pairs of complex conjugates ( $\lambda_1 = 0$ ;  $\lambda_\ell = \lambda_\ell^r + j\lambda_\ell^i$ ,  $\ell = 2, \dots, N$ ). It can be proved that (i)  $0 < \lambda_2^r \leq \dots \leq \lambda_N^r \leq 2$ , and (ii)  $|\lambda_\ell^i| \leq 1$ ,  $\forall \ell$ . The best propensity for synchronization is then ensured when both the ratio  $\lambda_N^r/\lambda_\ell^r$  and  $M \equiv \max_\ell \{|\lambda_\ell^i|\}$  are simultaneously made as small as possible.

In scale-free network models, the age of a node can be denoted by the time when it is being added to the network. The class of scale-free networks under study is generated from the Barabási–Albert model [82,83]. For comparison, a highly homogeneous random network with an arbitrary initial age ordering is considered, with the average degree being equal to that of the scale-free network. Figure 9 shows the variation of the synchronizability of the two networks versus the parameter  $\theta$ . For the random network, symmetric coupling makes the ratio  $\lambda_N^r/\lambda_2^r$  smallest, while for the scale-free model, the propensity for synchronization is better (or worse) when  $\theta \rightarrow -1$  (or  $\theta \rightarrow 1$ ). As for  $M$ , there are only very small differences between



**Synchronization Phenomena on Networks, Figure 9**

**a**  $\lambda_N^r$  and  $\lambda_2^r$ , **b**  $\lambda_N^r/\lambda_2^r$ , and **c**  $M$ , vs.  $\theta$ , for a scale-free network ( $m = 5$  and  $B = 0$ , solid curves) and random network (dashed curves) (after [81])

the scale-free and the random-network models. Thus, it is concluded that in scale-free networks, the network synchronizability is enhanced when the dominant coupling direction is from older to younger nodes [84].

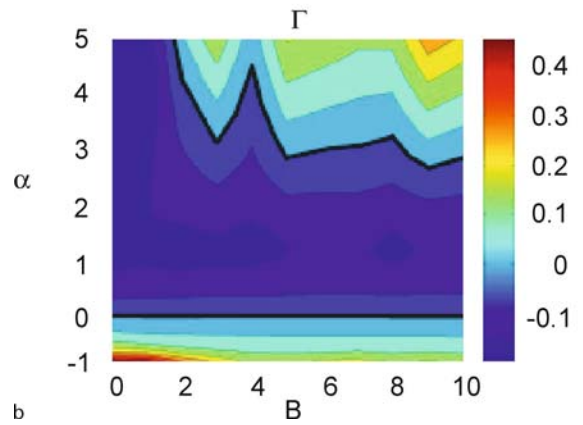
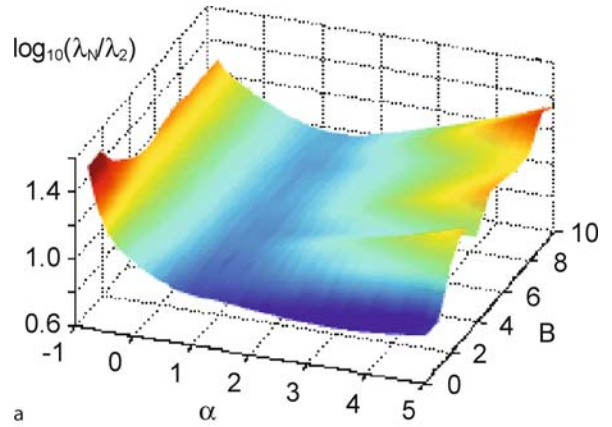
Taking the edge-weights into account, Chavez et al. [85,86] investigated the propensity for synchronization of some weighted complex networks, where the weight in an arbitrary edge,  $\ell_{ij}$ , is defined as its traffic load [87], which quantifies the traffic of shortest paths which make use of that edge. In this coupling pattern, the off-diagonal entries of the zero-row-sum coupling matrix  $G$  are

$$G_{ij} = -\frac{\ell_{ij}^\alpha}{\sum_{j \in \Delta_i} \ell_{ij}^\alpha} \ell_{ij}^\alpha, \quad (11)$$

where  $\alpha$  is a tunable parameter, and  $\ell_{ij}$  is the load of the edge connecting nodes  $i$  and  $j$ .

Although  $G$  is asymmetric for all  $\alpha$ , just like the MZK pattern, it can be proved that all its eigenvalues are non-negative reals with only one zero eigenvalue if the network is connected. The case of  $\alpha = 0$  corresponds to the best synchronizability condition for the MZK pattern. From Eq. (11), it can be seen that in the limit of  $\alpha = +\infty$  (or  $\alpha = -\infty$ ) only the edges with the largest (or smallest) loads  $\ell_{ij}$  are selected as the incoming edges for each node  $i$ . Therefore, this generates a network with at least  $N$  directed edges, which can be either connected or disconnected. In the connected (or disconnected) case, the ratio  $\lambda_N/\lambda_2$  will be equal to 2 (or  $+\infty$ ), thus yielding a very strong (or weak) condition for synchronization.

Figure 10a shows the logarithm of  $\lambda_N/\lambda_2$  in the parameter space  $(\alpha, B)$  for the above-discussed model [82,83]. Parameter  $B$  is used to regulate the heterogeneity of the

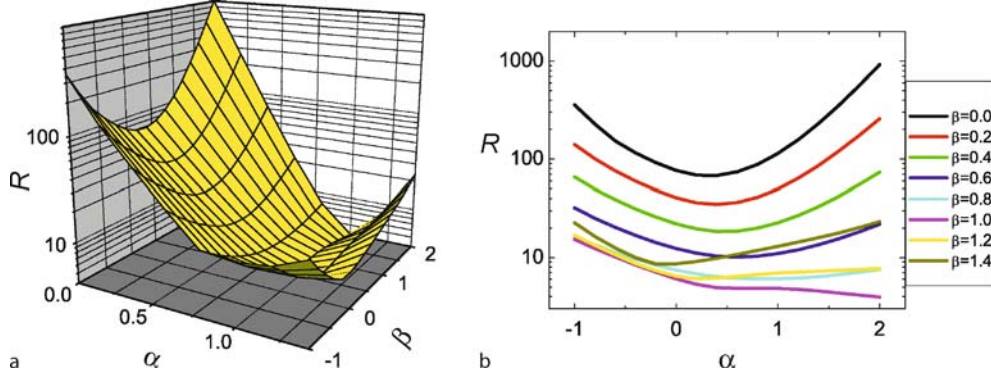


**Synchronization Phenomena on Networks, Figure 10**

**a**  $\lambda_N/\lambda_2$  for scale-free networks vs. the parameter space  $(\alpha, B)$ . **b**  $\Gamma = \log(\lambda_N/\lambda_2) - [\log(\lambda_N/\lambda_2)]_{\alpha=0}$  vs.  $(\alpha, B)$ . The domain with  $\Gamma < 0$  is outlined by the black contours drawn (after [85])

degree distribution. It can be observed that the surface of  $\lambda_N/\lambda_2$  has a prominent minimum when  $\alpha \simeq 1$  for all values of  $B$  above a given threshold  $B_c > 0$ , which means that the weighting procedure based on edge loads always enhances the network propensity for synchronization. The quantity  $\Gamma = \log(\lambda_N/\lambda_2) - [\log(\lambda_N/\lambda_2)]_{\alpha=0}$  shown in Fig. 10b may be used to quantify the synchronizability enhancement.

The coupling patterns proposed by both Hwang et al. [81] and Chavez et al. [85,86] can enhance the propensity for network synchronization. The former works well only for age-ordered networks, while the latter requires the knowledge of the load on each edge of the whole network. Therefore, a general coupling pattern using only local information would be very desirable. Based on the idea that different nodes should play different roles in a network, Zhao et al. [73] proposed a coupling pattern which



Synchronization Phenomena on Networks, Figure 11

a Eigenratio  $R$  in the parameter plane  $(\alpha, \beta)$ . b  $R$  vs.  $\alpha$  for different values of parameter  $\beta$  (after [73])

requires only the degrees of neighboring nodes. The coupling matrix  $G$  of this pattern is given by

$$G_{ij} = \begin{cases} -k_j^\alpha / S_i^\beta & \text{for } j \in \Lambda_i \\ S_i / S_i^\beta & \text{for } i = j \\ 0 & \text{otherwise,} \end{cases} \quad (12)$$

where  $S_i = \sum_{j \in \Lambda_i} k_j^\alpha$ . When  $\alpha = \beta = 0$ , this coupling pattern degenerates to the symmetric coupling pattern [45], where the case of  $\alpha = 0$  corresponds to the MZK pattern [78] and the case of  $\beta = 1$  is equivalent to the one introduced in [80] (see Eq. (15) in [80] for more details). Although this  $G$  is asymmetric for all  $\alpha$  with  $\beta \neq 0$ , it can also be proved that all its eigenvalues are non-negative reals with only one zero eigenvalue, if the network is connected. Figure 11 shows some simulation results. From the figure, it can be concluded that there is always some parameter region in which the eigenratio  $R$  is smaller than that of the symmetrically coupled case ( $\alpha = \beta = 0$ ) and that of the optimal case with the MZK pattern ( $\alpha = 0$  and  $\beta = 1$ ).

From the viewpoint of gradient fields, Wang et al. [88] also derived a coupling pattern that has the same configuration as Eq. (12) with  $\beta = 1$ .

### Dynamic Coupling Patterns

The coupling patterns discussed above are all based on a network having a fixed structure which remains unchanged throughout the synchronizing process.

Zhou et al. [89] investigated synchronization in a scale-free network of chaotic oscillators, where the coupling strength of a node develops adaptively according to the local synchronizing property between the node and its neighbors. In this coupling pattern, the off-diagonal en-

tries of the zero-row-sum coupling matrix  $G$  are

$$G_{ij} = -a_{ij} W_{ij}, \quad (13)$$

where  $W_{ij} > 0$  is the coupling strength from node  $j$  to node  $i$  if they are connected. Here, suppose that the strength between node  $i$  and all its  $k_i$  neighbors increases uniformly among the  $k_i$  connections, in order to suppress its difference  $\Delta_i$  from the mean activity of its neighbors; namely,

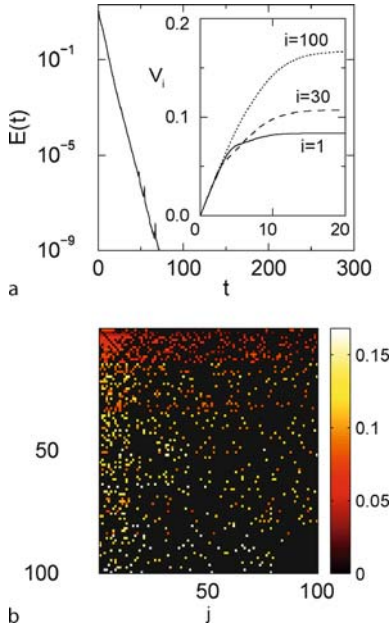
$$G_{ij}(t) = -a_{ij} V_i(t), \quad \dot{V}_i = \gamma \Delta_i / (1 + \Delta_i), \quad (14)$$

where  $\Delta_i = |\mathbf{H}(\mathbf{x}_i) - (1/k_i) \sum_j a_{ij} \mathbf{H}(\mathbf{x}_j)|$ , and  $\gamma > 0$  is the adaptation parameter. It is clear that, in this adaptive coupling scheme, the input weight ( $W_{ij} = V_i$ ) and the output weight ( $W_{ji} = V_j$ ) of node  $i$  are generally asymmetrical.

Next, synchronization of a network of coupled Rössler oscillators and a chaotic foodweb model on Barabási–Albert scale-free networks are considered, and two cases of unbounded and bounded stability zones are investigated, respectively. When the stability zone is unbounded, the transition to synchronization is shown in Fig. 12a. Starting from random initial conditions on the chaotic attractors, the local synchronization difference  $\Delta \gg 1$ , and the input weights of each node, both increase uniformly on the whole network, i.e.,  $W_{ij} = V_i(t) \approx \gamma t$  (Fig. 12a, inset). After a short period of time, the weights  $V_i$  of different nodes develop at different rates and then converge to different values  $\tilde{V}_i$ . The input weight is smaller on average for nodes with larger degrees  $k_i$  (Fig. 12b). Here, the synchronization error is measured by averaging all local errors over the nodes:  $E(t) = \langle |\mathbf{x}_i - \langle \mathbf{x}_i \rangle| \rangle$ .

The dependence of the input weight of a node on its degree follows a power law,

$$V(k) \sim k^{-\theta}, \quad (15)$$



**Synchronization Phenomena on Networks, Figure 12**

**a** Transition to synchronization in an adaptive network of Rössler oscillators, indicated by the (averaged) synchronization error  $E(t) = \langle |x_i - \langle x_i \rangle| \rangle$ . *Inset:* the input strength  $V_i(t)$  vs. time over three nodes. **b** The weighted coupling matrix  $\tilde{G}$  crystallized after the adaptation (for the foodweb model) (after [89])

with exponent  $\theta = 0.48 \pm 0.01$  for both oscillator models. Importantly, this scaling is also robust to the variation of network parameters, such as the minimal degree  $M$  (Fig. 13b), which should not be confused with the order parameter  $M$  elsewhere, the system size  $N$  (Fig. 13c), and the orders of magnitudes of the adaptation parameter  $\gamma$  (Fig. 13d).

When the stability zone is bounded, synchronization can always be achieved by the adaption mechanism of Eq. (14) if  $\gamma \leq \gamma_c$  for a threshold  $\gamma_c$  somewhat depending on  $N$  and the oscillator dynamics. The two resulting weighted networks display the same power-law behavior as in Eq. (15), but with different exponents:  $\theta = 0.54 \pm 0.01$  (Rössler oscillator) or  $\theta = 0.36 \pm 0.01$  (foodweb). The eigenratio  $R$  for the weighted networks, after the adaptation, and for the unweighted networks (symmetric coupling) is calculated as a function of  $N$  (Fig. 14a), and as a function of the ratio  $S_{\max}/S_{\min}$  (Fig. 14b), where  $S_{\max}$  and  $S_{\min}$  are the maximum and minimum intensities of the variable coupling strengths of the model. Clearly, this adaptive coupling scheme is more effective than symmetric coupling for network synchronization.

Huang [90] investigated another adaptive coupling pattern, in which a node is coupled with its neighbors

non-uniformly through different coupling strengths, and showed that they have better synchronizability than other networks with symmetric coupling patterns.

In all the coupling patterns discussed above, whether the coupling pattern is static or dynamic, only the coupling strength is tunable while the connectivity matrix always remains unchanged. However, as is intuitively clear, network synchronizability can also be significantly improved by evolving the graph topology giving rise to a time-varying connectivity matrix. This has been recently confirmed by Boccaletti et al. [91].

It has been shown [91] that to make a network synchronizable, either the coupling matrix  $G(t) = G$  remains unchanged, or if starting from an initial wiring condition  $G(0) = G_0$ , the coupling matrix  $G(t)$  commutes at any time with  $G_0$ , i. e.,  $G_0 G(t) = G(t) G_0, \forall t$ . At any time, a zero-row-sum symmetric commuting matrix  $G(t)$  can be constructed, as

$$G(t) = V \Lambda(t) V^T, \quad (16)$$

where  $V = \{v_i, \dots, v_N\}$  is an orthogonal matrix with columns being the eigenvectors of  $G_0$ , and  $\Lambda(t) = \text{diag}[0, \lambda_2(t), \dots, \lambda_N(t)]$  with  $\lambda_i(t) > 0, \forall i > 1$ . This set of matrices is referred to as the dissipative commuting set of  $G(0)$ . A condition to ensure the network synchronization will be stable is

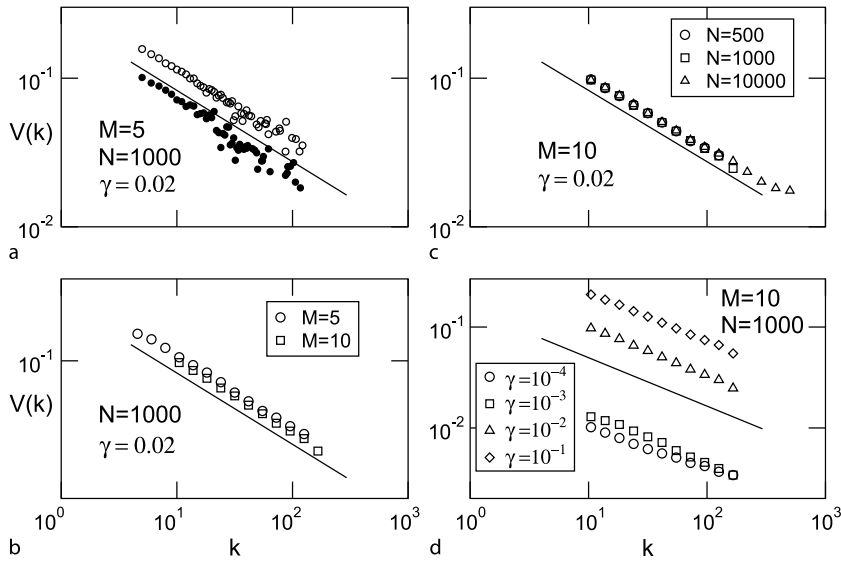
$$S_i = \lim_{T \rightarrow \infty} \frac{1}{T} \int_0^T \Lambda_{\max}(\sigma \lambda_i(t')) dt' < 0 \quad \forall i \neq 1, \quad (17)$$

where  $\Lambda_{\max}(\sigma \lambda_i)$  is the maximal transversal (conditional) Lyapunov exponent along the direction of the  $i$ th eigenvector, and  $S_i$  is its time average. Hence, it does not require  $\Lambda_{\max}(\sigma \lambda_i(t)) < 0$  at all times. One can even construct a commutative evolution such that at each time there exists one eigenvalue  $\lambda_i$  for which  $\Lambda_{\max}(\sigma \lambda_i(t)) > 0$ , and yet obtain a stable synchronization manifold. Thus, interestingly, synchronization in a dynamical network can be achieved even in the case where each individual commutative graph does not give rise to synchronized behavior.

### Modifications of Network Structures

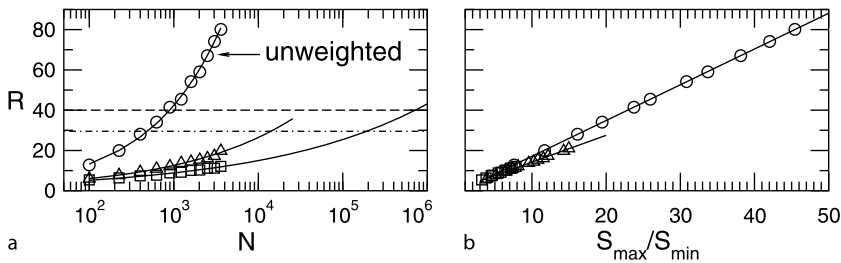
It is well known that the synchronizability of a dynamical network is determined simultaneously by the network coupling pattern, the dynamical characteristics of the oscillators on its nodes, and the network structure. In the above, several cases with variable coupling patterns have been discussed. For some real-world networks, however, the coupling pattern cannot be modified at will. Thus, if the dynamics of the oscillators are given and fixed, and





**Synchronization Phenomena on Networks, Figure 13**

Average input weight  $V(k)$  of nodes with degree  $k$  as a function of  $k$  for a network of Rössler oscillators (*empty circles*) and the food-web model (*filled circles*) (a), and its dependence on various parameters,  $M$  (b),  $N$  (c), and  $\gamma$  (d), where the  $M$  should not be confused with the order parameter  $M$  elsewhere (after [89])



**Synchronization Phenomena on Networks, Figure 14**

Eigenratio  $R$  as a function of  $N$  (a), and  $S_{\max}/S_{\min}$  (b). The networks are synchronizable if  $R < R_{\epsilon}$  in a, Rössler oscillators (*squares*),  $R_{\epsilon} = 40$  (*dashed curve*), foodweb model (*triangles*),  $R_{\epsilon} = 29$  (*dashed-dotted curve*) (after [89])

if the coupling patterns cannot be changed, then the only way to enhance the network synchronizability is to make a change to the network structure.

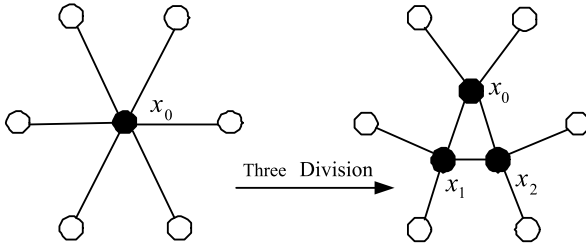
There are some effective techniques to enhance the network synchronizability by modifying the network structure, as further discussed below in the rest of this section.

### Reducing Maximal Betweenness

In scale-free networks, the average distance is often very short while the node-degree and node-betweenness distributions are both quite broad. The bottleneck for the network synchronizability seems to be the maximal node betweenness [57]. In order to reduce the node betweenness of the hubs, Zhao et al. [92] suggested a method of struc-

tural perturbations. Specifically, for a hub  $x_0$ ,  $m - 1$  auxiliary nodes, labeled as  $x_1, \dots, x_{m-1}$ , are added around it. These  $m$  nodes are fully connected together. Then, all the edges incident from  $x_0$  are re-distributed to all the nodes  $x_i$  (including  $x_0$  itself),  $i = 0, 1, \dots, m - 1$ . After this process, the betweenness of  $x_0$  is divided into  $m$  almost equal parts associating with these  $m$  nodes. This process is called  $m$ -division. A sketch map of a 3-division process on node  $x_0$  is shown in Fig. 15.

Due to the huge sizes of many real-life networks, it is usually impossible to obtain the node betweenness from a complex network. Fortunately, studies have shown that there exists a strongly positive correlation between the node-degree and the node-betweenness in Barabási-Albert networks and some other heterogeneous networks [87,93]. That is, a node with larger degree has

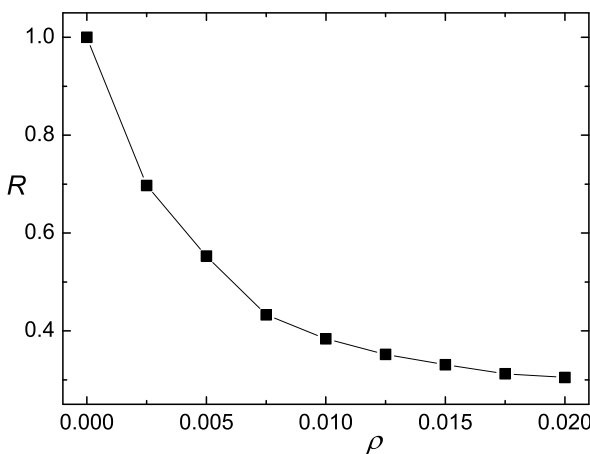


Synchronization Phenomena on Networks, Figure 15

Sketch map for the 3-division process on  $x_0$ . The solid circle on the left is the node  $x_0$  with degree 6. After the 3-division process, this  $x_0$  is divided into 3 nodes,  $x_0$ ,  $x_1$  and  $x_2$ , which are fully connected. The six edges incident from  $x_0$  are then re-distributed to all the three nodes (after [92])

higher node-betweenness statistically. Therefore, for practical reasons, it can be assumed that nodes with higher betweenness are those with larger degrees in Barabási–Albert networks.

To further explore how the structural perturbations affect the network synchronizability, the eigenratios before and after the  $m$ -division process were compared in [92] for a Barabási–Albert scale-free network with the coupling matrix being Laplacian. For use in the rest of the article, we define a characteristic value  $R = r'/r$ , in which  $r$  and  $r'$  are the eigenratios before and after the division, respectively. Figure 16 shows the correlation between  $R$  and the probability  $\rho$  of the divided nodes. It is clear that even the  $m$ -division of a tiny fraction of nodes can sharply enhance network synchronizability.



Synchronization Phenomena on Networks, Figure 16

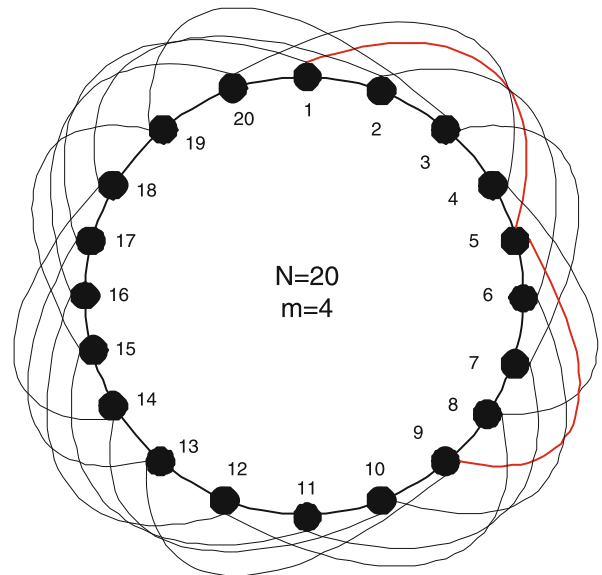
Behavior of value  $R$  vs. the fraction of divided nodes  $\rho$ . As the number of divided nodes increases,  $R$  is reduced, leading to better synchronization (after [92])

### Shortening the Average Distance

Zhou et al. [94] investigated the synchronizability of a network model named *crossed double cycles* (CDCs). They not only clarified the relationship between average distance and network synchronizability, but also provided a possible way to make a network more synchronizable.

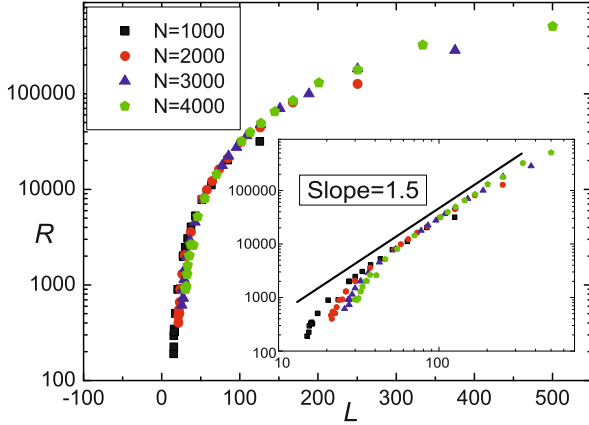
In the language of graph theory [95,96,97], a cycle  $C_N$  denotes a network consisting of  $N$  nodes (vertices)  $x_1, \dots, x_N$ . These  $N$  nodes are arranged in a ring, and the nearest two nodes are connected to each other. Thus,  $C_N$  has  $N$  edges connecting the nodes  $x_1x_2, x_2x_3, \dots, x_{N-1}x_N, x_Nx_1$ . The set of all such CDCs, denoted by  $G(N, m)$ , can be constructed by adding two edges, called *crossed edges*, to each node in  $C_N$ . The two nodes connecting by a crossed edge have distance  $m$  within  $C_N$ . For example, the network  $G(N, 3)$  can be constructed from  $C_N$  by connecting  $x_1x_4, x_2x_5, \dots, x_{N-1}x_2, x_Nx_3$  together. A sketch map of  $G(20, 4)$  is shown in Fig. 17 for illustration.

Figure 18 shows how the average distance  $L$  affects the network synchronizability (measured by the characteristic value  $R$ ). It is clear that the network synchronizability is very sensitive to the average distance: as  $L$  increases,  $R$  sharply spans more than three magnitudes. And the network synchronizability is remarkably enhanced by reducing  $L$ . When the crossed length  $m$  is not too small or too large (compared to  $N$ ), networks with the same average distance have approximately the same synchronizability,



Synchronization Phenomena on Networks, Figure 17

Sketch map of  $G(20, 4)$  (after [94])



**Synchronization Phenomena on Networks, Figure 18**

Characteristic value  $R$  vs. average distance  $L$  of CDCs. The black squares, red circles, blue triangles and green pentagons represent the cases of  $N = 1000, 2000, 3000$  and  $4000$ , respectively. The inset shows the same data in log-log plot, indicating that the characteristic value  $R$  approximately obeys a power-law form  $R \sim L^{1.5}$ . The solid line has slope 1.5, for comparison (after [94])

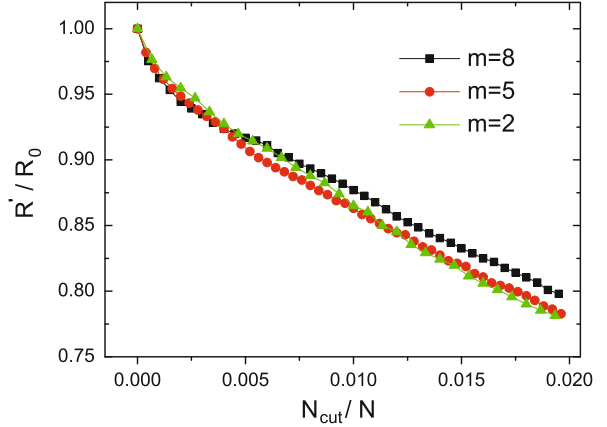
regardless of the network sizes. More interestingly, the numerical results show that the characteristic value  $R$  approximately obeys a power-law form, as  $R \sim L^{1.5}$  (inset of Fig. 18).

### Decoupling Nodes by Removing Heavily-Loaded Edges

In the synchronization process, not only hubs may be the bottlenecks but some edges with large loads may also limit the network synchronizability. Yin et al. [99] found that a scale-free network can become more synchronizable after some of its heavily-loaded edges have been removed. To reduce the computational cost, they used local information to approximately rank the edges, according to the values of  $k_i \times k_j$ , where  $i$  and  $j$  denote two adjacent nodes connected by an edge. Subsequently, at each time step, an edge with the highest rank is removed, i.e., the two nodes are decoupled at both sides of their connecting heavily-loaded edge. After this operation, the characteristic value is decreased, as shown by Fig. 19.

### Designing the Output Function

Very recently, the relationship between graph theory and network synchronizability received some special attention [100]. For example, Duan et al. [101,102,103] found that for networks with disconnected complementary graphs, adding edges will often increase their synchronizability. The complementary graph of a given graph  $G$  is



**Synchronization Phenomena on Networks, Figure 19**

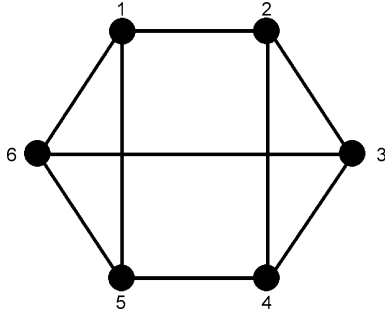
Changes of the synchronizability as a function of the proportion of cut edges  $N_{\text{cut}}/N$  for different values of the average distance (after [99])

defined to be the graph consisting of all the nodes of  $G$  and all the edges that are not in  $G$ .

In addition, they found [101,102] that when the couplings between nodes are symmetric, an unbounded synchronized region is always easier to analyze than a bounded synchronized region (see Sect. “Basic Concepts of Network Synchronization” to recall their definitions). Therefore, to effectively enhance network synchronizability, they presented a design method for the output function (i.e.,  $H$  in network (1), or the inner linking matrix in the linear coupling case), such that the resultant network has an unbounded synchronized region, for the case where the synchronous state is an equilibrium of the network.

If the synchronous state is an equilibrium, then both  $DF(s(t))$  and  $DH(s(t))$  in network (1), as discussed in Sect. “Enhancing Network Synchronizability” (part B), reduce to constant matrices, denoted by  $F$  and  $H$ , respectively. The synchronized region is the stability region of the matrix pencil  $F + \alpha H$  with respect to parameter  $\alpha$ . It can be proved that there exists a matrix  $H$  of rank 1 (meaning that only one component in each state vector is used for coupling), such that the stability region is unbounded. The method for obtaining the desired output function is outlined below: first, take a column vector  $b$  such that  $(F, b)$  is stabilizable [104]; then, find a matrix  $P = P^T$  such that  $FP + PF^T - 2bb^T < 0$ ; consequently, taking  $k = b^T P^{-1}$  leads to the stability of  $F - \alpha bk$  for all  $\alpha$  in the unbounded region; finally,  $H = bk$  is the matrix to be found.

For illustration, synchronization of a simple 6-node network (shown in Fig. 20) is studied, where each node



**Synchronization Phenomena on Networks, Figure 20**  
A network of 6 nodes (after [101])

is located with a third-order smooth Chua's circuit [105]. At first, arbitrarily take the output function

$$H = \begin{pmatrix} 0.8348 & 9.6619 & 2.6591 \\ 0.1002 & 0.0694 & 0.1005 \\ -0.3254 & -8.5837 & -0.9042 \end{pmatrix}. \quad (18)$$

But the network does not synchronize. The states of node 1 are shown in Fig. 21a. Then, let  $b = (0, 0, 1)^T$  and  $k = (0.0708, -0.15590, 0.4296)$ , and then set  $H = bk$ , so synchronization is achieved as guaranteed by the theory. Figure 21b shows that the states of node 1 quickly reach the equilibrium.

### Future Research Outlook

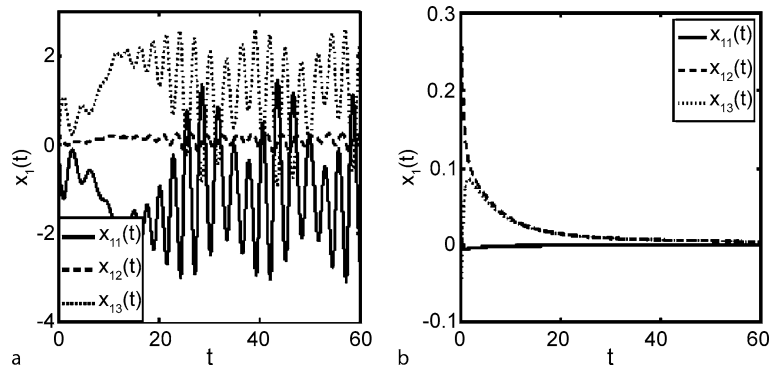
Complex network synchronization is a rapidly growing subject attracting increasing attention from various fields of physics, engineering, mathematics, and biology alike. Despite the current great advances and progress, there are still many important open questions.

In the studies of static coupling, Nishikawa and Motter [48] once pointed out that optimal global synchro-

nizability, with eigenratio being equal to 1, can be obtained from a directed network structure without loops. Even if adding one loop of length 2 (in a directed network, two opposite edges between node  $i$  and node  $j$  can be considered as a loop of length 2), the eigenratio will be doubled [73,85]. Another scenario is shown by extending the conclusion in [48] to the case of non-identical oscillators [106]. Some further works in this direction will be helpful for in-depth understanding about the role of loops in network synchronization.

In the studies of dynamic coupling, the cost of coupling has not been taken into account. However, cost is usually very significant in some self-driven systems (for example, in wireless sensor networks [107,108] and in distributed autonomous robotic systems [109]). For each node to report its current state to the neighbors (or to detect the states of all its neighbors) requires a certain amount of power, while the total power assigned to each node is often limited, even if such communications are possible. Yet, as found in collective behaviors of biological swarms, a few effective leaders can well organize the whole population [110]. And a recent study has pointed out that partial coupling is more than enough to keep the coherence of self-propelled particles [111]. Therefore, it is very natural to expect to synchronize a complex network with a very low cost, which is an important issue for further investigation.

Very recently, there are some attempts at detecting the network structures with the help of the synchronization phenomenon on complex networks [112,113,114]: to discover the hierarchical community structure by the dynamic time scales of the network synchronization process [112,113], or to infer the complete connectivity of a network from its stable response dynamics [114], etc. These seem quite useful for optimal network design, analy-



**Synchronization Phenomena on Networks, Figure 21**  
States of node 1 (after [101])

sis, and utilization in general, therefore should be pursued with special efforts.

Similar to the aforementioned open questions, many theoretically attractive and practically important problems about various aspects of synchronization on complex networks can be posted and described. As the network research further evolves in different fields, many new dynamical phenomena and analytic issues will also emerge. Importance notwithstanding, the subject of “Synchronization Phenomenon of Networks” will continue to prove itself an theoretically interesting and technically challenging subject for scientific research in the years to come.

### Acknowledgments

The authors thank the American Physics Society and the Elsevier Publisher for their permission to use some of their published simulation figures, which have all been referenced (in the figure captions). The authors also thank Dr. Jian-Guo Liu for his helpful comments and suggestions. G.R.C. was supported by the NSFC/RGC joint research scheme under grant N-CityU107/07. B.H.W. was supported by the NNSFC under Grants 10472116 and A0524701, and the Specialized Program under the Presidential Funds of CAS. T.Z. was supported by NNSFC under Grants 10635040 and 70471033.

### Bibliography

- Albert R, Barabási AL (2002) *Rev Mod Phys* 74:47
- Dorogovtsev SN, Mendes JFF (2002) *Adv Phys* 51:1079
- Newman MEJ (2003) *SIAM Rev* 45:167
- Boccaletti S, Latora V, Moreno Y, Chaves M, Hwang DU (2006) *Phys Rep* 424:175
- da F Costa L, Rodrigues FA, Travieso G, Boas PRV (2007) *Adv Phys* 56:167
- Pastor-Satorras R, Vespignani A (2001) *Phys Rev Lett* 86:3200
- Zhu CP, Xiong SJ, Tian YJ, Li N, Jiang KS (2004) *Phys Rev Lett* 92:218702
- Zhou T, Yan G, Wang BH (2005) *Phys Rev E* 71:046141
- Zhou T, Fu ZQ, Wang BH (2006) *Prog Nat Sci* 16:452
- Zhou T, Liu JG, Bai WJ, Chen GR, Wang BH (2006) *Phys Rev E* 74:056109
- Motter AE, Lai YC (2002) *Phys Rev E* 66:065102
- Goh KI, Lee DS, Kahng B, Kim D (2003) *Phys Rev Lett* 91:148701
- Motter AE (2004) *Phys Rev Lett* 93:098701
- Zhou T, Wang BH (2005) *Chin Phys Lett* 22:1072
- Galstyan A, Cohen P (2007) *Phys Rev E* 75:036109
- Guimerá R, Díaz-Guilera A, Vega-Redondo F, Cabrales A, Arenas A (2002) *Phys Rev Lett* 89:248701
- Tadić B, Thurner S, Rodgers GJ (2004) *Phys Rev E* 69:036102
- Zhao L, Lai YC, Park K, Ye N (2005) *Phys Rev E* 71:026125
- Yan G, Zhou T, Hu B, Fu ZQ, Wang BH (2006) *Phys Rev E* 73:046108
- Wang BH, Zhou T (2007) *J Korean Phys Soc* 50:134
- Watts DJ, Strogatz SH (1998) *Nature* 393:440
- Barabási AL, Albert R (1999) *Science* 286:509
- Strogatz SH, Stewart I (1993) *Sci Am* 269:102
- Gray CM (1994) *J Comput Neurosci* 1:11
- Néda Z, Ravasz E, Vicsek T, Barabási AL (2000) *Phys Rev E* 61:6987
- Glass L (2001) *Nature* 410:277
- Chen G, Dong X (1998) *From Chaos Order*. World Scientific, Singapore
- Erdős P, Rényi A (1960) *Pub Math Ins Hung Acad Sci* 5:17
- Wang XF (2002) *Int J Bifurc Chaos* 12:885
- Heidelberg R (2007) *Nature* 450:625
- Heagy JF, Carroll TL, Pecora LM (1994) *Phys Rev E* 50:1874
- Stefański A, Perlikowski P, Kapitaniak T (2007) *Phys Rev E* 75:016210
- Wu CW, Chua LO (1995) *IEEE Trans Circuits Syst I* 42:430
- Jost J, Joy MP (2001) *Phys Rev E* 65:016201
- Gade PM (1996) *Phys Rev E* 54:64
- Manrubia SC, Mikhailov AS (1999) *Phys Rev E* 60:1579
- Gade PM, Hu CK (1999) *Phys Rev E* 60:4966
- Gade PM, Hu CK (2000) *Phys Rev E* 62:6409
- Lago-Fernández LF, Huerta R, Corbacho F, Sigüenza JA (2000) *Phys Rev Lett* 84:2758
- Wang XF, Chen G (2002) *Int J Bifurc Chaos Appl Sci Eng* 12:187
- Wang XF, Chen G (2002) *IEEE Trans Circuits Syst I* 49:54
- Barahona M, Pecora LM (2002) *Phys Rev Lett* 89:054101
- Jiang PQ, Wang BH, Bu SL, Xia QH, Luo XS (2004) *Int J Mod Phys B* 18:2674
- Lind PG, Gallas JAC, Herrmann HJ (2004) *Phys Rev E* 70:056207
- Pecora LM, Carroll TL (1998) *Phys Rev Lett* 80:2109
- Hu G, Yang J, Liu W (1998) *Phys Rev E* 58:4440
- Pecora LM, Barahona M (2005) *Chaos Complex Lett* 1:61
- Nishikawa T, Motter AE (2006) *Phys Rev E* 73:065106
- Kuramoto Y (1975) In: Araki H (ed) *International Symposium on Mathematical Problems in Theoretical Physics*. Lecture Notes in Physics, vol 30. Springer, New York
- Kuramoto Y (1984) *Chemical Oscillations, Wave and Turbulence*. Springer, Berlin
- Kuramoto Y, Nishikawa I (1987) *J Stat Phys* 49:569
- Pikovsky A (2001) *Synchronization*. Cambridge University Press, Cambridge
- Acebrón JA, Bonilla LL, Vicente CJP, Ritort F, Spigler R (2005) *Rev Mod Phys* 77:137
- Nishikawa T, Motter AE, Lai YC, Hoppensteadt FC (2003) *Phys Rev Lett* 91:014101
- Newman MEJ, Strogatz SH, Watts DJ (2001) *Phys Rev E* 64:026118
- Dorogovtsev SN, Mendes JFF (2000) *Phys Rev E* 62:1842
- Hong H, Kim BJ, Choi MY, Park H (2004) *Phys Rev E* 69:067105
- Freeman L (1979) *Soc Netw* 1:215
- Newman MEJ (2001) *Phys Rev E* 64:016132
- Zhou T, Liu JG, Wang BH (2006) *Chin Phys Lett* 23:2327
- Fan ZP (2006) *Complex Networks: From Topology to Dynamics*. Ph D Thesis, City University of Hong Kong
- Zhao M, Zhou T, Wang BH, Yan G, Yang HJ, Bai WJ (2006) *Physica A* 371:773
- Maslv S, Sneppen K (2002) *Science* 296:910
- Kim BJ (2004) *Phys Rev E* 69:045101(R)
- McGraw PN, Menzinger M (2005) *Phys Rev E* 72:015101
- Gómez-Gardeñes J, Moreno Y, Arenas A (2007) *Phys Rev Lett* 98:034101



67. Wu X, Wang BH, Zhou T, Wang WX, Zhao M, Yang HJ (2006) *Chin Phys Lett* 23:1046
68. Holme P, Kim BJ (2002) *Phys Rev E* 65: 026107
69. Newman MEJ (2002) *Phys Rev Lett* 89:208701
70. di Bernardo M, Garofalo F, Sorrentino F (2007) *Int J Bifurc Chaos* 17:3499
71. Sorrentino F, di Bernardo M, Cuéllar GH, Boccaletti S (2006) *Physica D* 224:123
72. Chavez M, Hwang DU, Martinerie J, Boccaletti S (2006) *Phys Rev E* 74:066107
73. Zhao M, Zhou T, Wang BH, Ou Q, Ren J (2006) *Eru Phys J B* 53:375
74. Huang L, Park K, Lai YC, Yang L, Yang K (2006) *Phys Rev Lett* 97:164101
75. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D (2004) *Proc Natl Acad Sci USA* 101:2658
76. Zhou T, Zhao M, Chen G, Yan G, Wang BH (2007) *Phys Lett A* 368:431
77. Donetti L, Hurtado PI, Muñoz MA (2005) *Phys Rev Lett* 95:188701
78. Motter AE, Zhou C, Kurths J (2005) *Phys Rev E* 71:016116
79. Motter AE, Zhou C, Kurths J (2005) *Europhys Lett* 69:334
80. Motter AE, Zhou C, Kurths J (2005) *AIP Conf Proc* 776:201
81. Hwang DU, Chavez M, Amann A, Boccaletti S (2005) *Phys Rev Lett* 94:138701
82. Dorogovtsev SN, Mendes JFF, Samukhin AN (2000) *Phys Rev Lett* 85:4633
83. Krapivsky PL, Redner S (2001) *Phys Rev E* 63:066123
84. Zou Y, Zhu J, Chen G (2006) *Phys Rev E* 74:046107
85. Chavez M, Hwang DU, Amann A, Hentschel HGE, Boccaletti S (2005) *Phys Rev Lett* 94:218701
86. Chavez M, Hwang DU, Amann A, Boccaletti S (2006) *Chaos* 16:015106
87. Goh KI, Kahng B, Kim D (2001) *Phys Rev Lett* 87:278701
88. Wang X, Lai YC, Lai CH (2007) *Phys Rev E* 75:056205
89. Zhou C, Kurths J (2006) *Phys Rev Lett* 96:164102
90. Huang D (2006) *Phys Rev E* 74:046208
91. Boccaletti S, Hwang DU, Chavez M, Amann A, Kurths J, Pecora LM (2006) *Phys Rev E* 74:016102
92. Zhao M, Zhou T, Wang BH, Wang WX (2005) *Phys Rev E* 72:057102
93. Barthélemy M (2004) *Eur Phys J B* 38:163
94. Zhou T, Zhao M, Wang BH (2006) *Phys Rev E* 73:037101
95. Bondy JA, Murty USR (1976) *Graph Theory with Applications*. MacMillan, London
96. Bollobás B (1998) *Modern Graph Theory*. Springer, New York
97. Xu JM (2003) *Theory and Application of Graphs*. Kluwer, Dordrecht
98. Newman MEJ, Watts DJ (1999) *Phys Rev E* 60:7332
99. Yin CY, Wang WX, Chen G, Wang BH (2006) *Phys Rev E* 74:047102
100. Comellas F, Gago S (2007) *J Phys A Math Theor* 40:4483
101. Duan Z, Chen G, Huang L (2007) *Phys Rev E* 76:056103
102. Duan Z, Chen G, Huang L (2007) *Phys Lett A* 372:3741
103. Duan Z, Liu C, Chen G (2008) *Physica D* 237:1006
104. Friedland B (1986) *Control System Design*. McGraw-Hill, New York
105. Tsuneda A (2005) *Int J Bifurc Chaos* 15:1
106. Um J, Han SG, Kim BJ, Lee SI (2008) (unpublished)
107. Akyildiz IF, Su W, Sankarasubramaniam Y, Cayirci E (2002) *Comput Netw* 38:393
108. Oqren P, Fiorelli E, Leonard NE (2004) *IEEE Trans Automat Contr* 49:1292
109. Arai T, Pagello E, Parker LE (2002) *IEEE Trans Robot Automat* 18:655
110. Couzin LD, Krause J, Franks NR, Levin SA (2005) *Nature* 433:513
111. Zhang HT, Chen M, Zhou T (2007) [arXiv:0707.3402](https://arxiv.org/abs/0707.3402)
112. Arenas A, Díaz-Guilera A, Pérez-Vicente CJ (2006) *Phys Rev Lett* 96:114102
113. Boccaletti S, Ivanchenko M, Latora V, Pluchino A, Rapisarda A (2007) *Phys Rev E* 75:045102(R)
114. Timme M (2007) *Phys Rev Lett* 98:224101

---

## Synergetics: Basic Concepts

HERMANN HAKEN

Institut für Theoretische Physik, Universität Stuttgart,  
Stuttgart, Germany

### Article Outline

[Glossary](#)

[The Role of Synergetics in Science](#)

[The Laser Paradigm](#)

[The Hierarchical Structure of Synergetics](#)

[Basic Equations](#)

[Method of Solution](#)

[Quantum Theoretical Formulation](#)

[Regular Spatial and Spatio-Temporal Patterns](#)

[A Further Mathematical Tool: Shannon Information  
and the Maximum \(Information\) Entropy Principle](#)

[Phenomenological Synergetics](#)

[Semantic Synergetics](#)

[Some Selected Examples](#)

[History and Relations to Other Fields](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Synergetics** Science of cooperation.

**Pattern** A pattern is essentially an arrangement. It is characterized by the order of the elements of which it is made rather than by the intrinsic nature of these elements (Norbert Wiener).

**Self-organization** Formation of spatio-temporal patterns (structures) and/or performance of functions without an “ordering hand”.

**State vector** Set of time- or time-independent variables that characterize the state of a system.

**Evolution equations** Determine the temporal evolution of the state vector. May be deterministic, stochastic or both.

**Control parameter** One or a set of (mostly externally) fixed parameters in the evolution equations.

**Spectrum** Set of eigenvalues belonging to linear stability equations with boundary conditions.

**Stability of a system** System returns after a (small) perturbation of its state vector into original state.

**Instability** Loss of stability.

**Order parameters** Collective variables that determine the macroscopic behavior of systems.

**Slaving principle** A general theorem that allows the reduction of the variables of a system to order parameters (close to instability).

**Trajectory** Smooth curve  $q(t)$  of solution of evolution equation in  $q$ -space.

**Attractor** Region in the state vector space (" $q$ -space") to which all neighboring states are attracted in the course of time.

**Fixed point, stable** Point in  $q$  space to which all neighboring trajectories converge in course of time.

**Limit cycle, stable** A closed trajectory to which all neighboring trajectories converge.

**Probability distribution function** Function that determines the probability of a random variable  $r$  to have fixed value  $r = r_0$ .

**Fokker Planck equation** Evolution equation for probability density function, based on drift and diffusion.

**Normal form** Especially simple polynomial expression that still captures the essential features, e. g. of the right hand side of deterministic evolution equations.

**Schrödinger picture of quantum mechanics** In it operators are time-independent, while the wave-function ("state vector") is time-dependent and determined by the Schrödinger equation.

**Heisenberg picture in quantum mechanics** The state vector is time-independent, while the operators are time-dependent and determined by Heisenberg equations of motion.

**Fluctuating forces** Stochastic (random) forces appearing in evolution equations.

**Quantum classical correspondence** Establishes relation between quantum mechanical density matrix and classical quasi-probability distribution.

**Symmetry** Invariance of a system against specific transformations (e. g. mirror symmetry).

**Group** Set of elements with specific multiplication rules (axioms).

**Dynamical system** System whose state vector changes in the course of time deterministically.

**Langevin equation** Originally: evolution equation for velocity of a Brownian particle subject to damping and fluctuating force.

**Generalized Langevin equation** General evolution equations that contain both a deterministic and a stochastic part ("fluctuating forces").

**Hamilton operator** Classical Hamilton function, in which variables, e. g. position  $x$  and momentum  $p$ , are replaced by quantum mechanical operators.

**Spatial coordinate (vector  $x$ )** in one, two or three dimensions.

$\Delta$  Laplace operator (in 1,2 or 3 dimensions) .

$\nabla$  Vector  $\left( \frac{d}{dx_1}, \frac{d}{dx_2}, \frac{d}{dx_3} \right)$  in 1,2 or 3 dimensions .

## The Role of Synergetics in Science

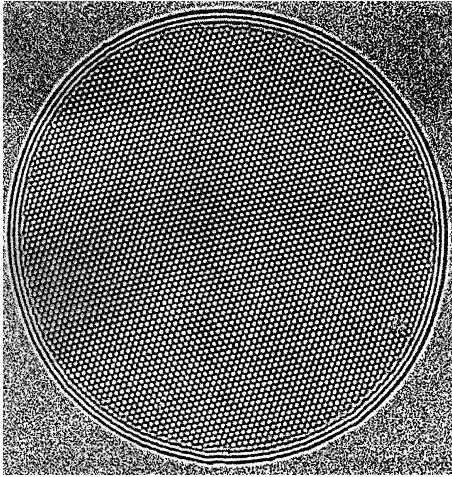
In science, we may essentially distinguish between two trends:

1. The accumulation of knowledge
2. Information reduction in the sense of finding general principles, common features.

In physics, such unifying approaches are well known: the unification of magnetism, electricity and, later on, weak and other interactions leading eventually to a unified field theory. General relativity unifies concepts of space, time and gravitation. While these unifications take place at a fundamental level, one may ask whether it is worthwhile to look also for unifications at say more macroscopic or phenomenological levels. One example is thermodynamics, another the theory of phase transitions of systems in thermal equilibrium by means of the renormalization group approach, or the concept of fractals, etc.

The main goal of Synergetics is the search for unifying principles for systems that are composed of many individual parts or components, and that may show the phenomenon of self-organization, i. e. the spontaneous formation of spatial, temporal, spatial-temporal or functional structures. The systems under discussion are, in the widest sense of the word, open physical systems whose states are maintained by an in- and outflux of energy, matter and/or information. A typical and well known example is that of a fluid in a pan that is uniformly heated from below. When the temperature difference between the lower and upper surface exceeds a critical value, the formerly homogeneous fluid develops roll or hexagonal patterns in which the fluid moves in a specific manner (Fig. 1).

As it turned out, the general principles originally elaborated in physics, can also be applied to many other systems, such as in biology, economy, ecology, sociology,



**Synergetics: Basic Concepts, Figure 1**

Hexagonal pattern of a fluid (liquid helium) uniformly heated from below [12]

management theory, psychology etc. In spite of the great variety of the individual systems with their components quite different in nature, such principles apply to large classes of phenomena. This is achieved by restricting the study to situations where the systems undergo qualitative changes at macroscopic scales. Here macroscopic means “with time and length scales large compared to those of the individual components”.

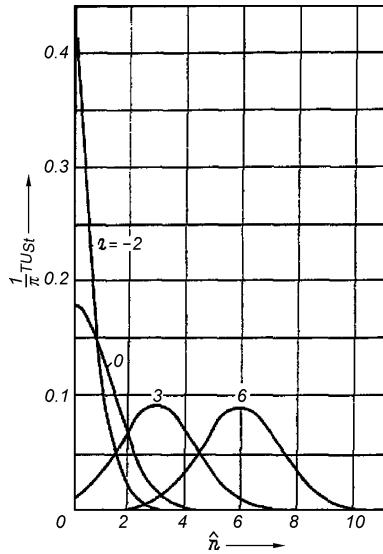
This leads to the definition of Synergetics as given in the preamble of the Springer Series in Synergetics: “An ever increasing number of scientific disciplines deal with complex systems. These are systems that are composed of many parts which interact with one another in a more or less complicated manner. One of the most striking features of many such systems is their ability to spontaneously form spatial or temporal structures. A great variety of these structures are found, in both the inanimate and the living world. In the inanimate world of physics and chemistry, examples include the growth of crystals, coherent oscillations of laser light, and the spiral structures formed in fluids and chemical reactions. In biology we encounter the growth of plants and animals (morphogenesis) and the evolution of species. In medicine we observe, for instance, the electromagnetic activity of the brain with its pronounced spatio-temporal structures. Psychology deals with characteristic features of human behavior ranging from simple pattern recognition tasks to complex patterns of social behavior. Examples from sociology include the formation of public opinion and cooperation or competition between social groups.”

In recent decades, it has become increasingly evident

that all these seemingly quite different kinds of structure formation have a number of important features in common. The task of studying analogies as well as differences between structure formation in these different fields has proved to be an ambitious but highly rewarding endeavor. The Springer Series in Synergetics provides a forum for interdisciplinary research and discussions on this fascinating new scientific challenge. It deals with both experimental and theoretical aspects. The scientific community and the interested layman are becoming ever more conscious of concepts such as self-organization, instabilities, deterministic chaos, nonlinearity, dynamical systems, stochastic processes, and complexity. All of these concepts are facets of a field that tackles complex systems, namely Synergetics.

### The Laser Paradigm

This example elucidates central concepts used in Synergetics in a qualitative fashion. An example for the laser device (an acronym for light amplification by stimulated emission of radiation, originally called optical maser [121]) is the gas laser in which gas atoms are enclosed in a tube at the end-faces of which mirrors are mounted. The mirrors serve the purpose of reflecting light running in axial direction sufficiently often so that the corresponding light wave stays for an extended period in this device and can interact intensely with the atoms. The atoms are excited from the outside, e. g. by a pump light source. After having been excited, each atom can spontaneously emit a light wave track. In the usual case of a *lamp*, these wave tracks are emitted independently of each other and the amplitudes are Gaussian distributed. When the pump intensity is increased beyond a critical value, the present state gives way to a single wave with *stable amplitude* on which small *amplitude fluctuations* and *phase diffusion* are superimposed [53]. The pump intensity serves as *control parameter*. At its critical value, the old state becomes *unstable*. The emerging coherent wave acts as *order parameter* that via stimulated emission forces the electrons of the gas molecules to emit light waves in a coherent fashion. This action of the order parameter on the individual parts of the system is called *slaving principle*. If the pump power is increased further, more instabilities can appear, and a variety of temporal but also spatio-temporal patterns of light waves may appear, such as laser light chaos [55] or ultrashort laser pulses. The first laser threshold shows the typical features of a phase transition of a system in thermal equilibrium, namely critical slowing down, critical fluctuations and symmetry breaking [25,46,53,58,119], as well as the emergence of a c-number amplitude of the quantized light field (Fig. 2).



**Synergetics: Basic Concepts, Figure 2**

The stationary distribution function of the laser light intensity as a function of the normalized intensity  $\hat{h}$ . The individual curves refer to different normalized pump power values  $a$ , where  $a < 0$  below threshold,  $a = 0$  at threshold,  $a > 0$  above threshold (after [116])

## The Hierarchical Structure of Synergetics

Before I discuss the mathematical approach in detail and to provide the ground for farther reaching applications, I hint at the three levels of Synergetics:

1. *The microscopic theory*, based either on microscopic equations, such as in the laser example, those of quantum mechanics and quantum field theory, or in biology on mathematical models on the behavior of individual parts of a system. At this level, concepts, such as order parameters and enslavement (cf. Sect. “[The Laser Paradigm](#)”), can be mathematically derived.
2. *Phenomenological Synergetics* directly starts from concepts, such as order parameters and enslavement, which then may be cast into mathematical relations.
3. *Semantic Synergetics* deals with cases where a mathematical formulation is (at present or in principle) not possible, but still formulations using concepts and relationships unearthed in Synergetics are applicable.

A general goal of Synergetics consists in elaborating relationships between levels 1, 2, 3.

In the present article I will mainly focus my attention on the mathematical formulation dealing with 1. and 2.

## Basic Equations

The basic equations are classical or quantum mechanical evolution equations, in which the temporal evolution of the microscopic quantities under consideration is described by ordinary or partial differential equations. Since the systems are open, the inputs and outputs of energy, matter and/or information must be taken care of, which, quite often, appears in the form of coupling to heat baths in the sense of thermodynamics. In open systems, these heat baths must be kept at different temperatures, in order to maintain the non-equilibrium state of the system. The heat bath variables can be eliminated which gives rise to differential equations which contain “pumping” and “damping” terms as well as fluctuating (stochastic) forces. In the case of quantum mechanical equations the stochastic forces are operators. With the inclusion of stochastic forces, the classical or quantum mechanical equations acquire the character of stochastic differential equations which may be called “generalized Langevin equations”.

Depending on the definition of the random forces, we may distinguish between the  $\hat{I}$  to, the Statonovich and the Klimontovich approach [62,72,134]. As is well known in statistical physics, Langevin equations can be converted into equations for distribution functions, such as e.g. the Fokker–Planck equation. A further approach, mainly used in quantum mechanics, but also in models on sociodynamics, is the master equation.

In order not to overload this article, I will focus my attention on the treatment of evolution equations.

This approach seems to be particularly suited for the treatment of phase transition- like phenomena, i.e. the transitions between qualitatively different states of a system. If noise is neglected and transients are not treated, these transitions are called bifurcations [5,22,49,71,79,84,91].

At the microscopic level the systems are described by a state vector  $q$  with components  $q_1, \dots, q_n$  which may also be space dependent,  $q_j = q_j(x, t)$ , where  $x$  is a one, two or three dimensional vector. The time dependence is described by evolution equations of the form of a vector equation.

$$\dot{q} = N(q, \nabla, \alpha) + F(q, \nabla, \alpha). \quad (1)$$

The dot  $\dot{\phantom{x}}$  means time-derivative.  $N$  is a vector valued function that depends on  $q$  in a nonlinear fashion.  $\nabla$  indicates spatial derivatives (of any order) or non-local integrations e.g. of the form

$$\int K(x, x') q(x') dx' \quad (2)$$

where  $K$  is a matrix.

$\alpha$  represents a set of fixed control parameters. If not otherwise stated, we explicitly treat only one control parameter. Equation (1) must be supplemented by appropriate boundary and initial conditions.  $F$  is a vector valued stochastic function of time with vanishing mean.

## Method of Solution

We assume that for a certain control parameter value  $\alpha_0$  the state vector as solution of Eq. (1) is known,  $q = q_0$ . The following cases have been considered, see e. g. [62]:

- $q_0$  is a stable fixed point (Section “Instability of a Fixed Point”)
- $q_0$  is a stable limit cycle (Section “Instability of a Limit Cycle,  $q_0(t)$  [62]”)
- $q_0$  is a stable  $n$ -dimensional torus. (Section “Instability of Tori [62]”)

Now the control parameter value is changed and the stability of the system is checked by means of linear stability analysis [52].

**Instability of a Fixed Point** We first elucidate our general procedure by means of the instability of an originally stable fixed point. This procedure differs from the classical approach of bifurcation theory [90,124] in two important aspects:

- The role of the fluctuating forces is fully taken into account in order to be able to make contact with the theory of phase transitions in the Landau sense [86].
- The approach covers the surrounding of the fixed point in order to deal with relaxation processes towards the newly evolving stable states.

The hypothesis

$$q(t) = q_0 + W(t) \quad (3)$$

is inserted into (1) and the Eq. (1) with  $F \equiv 0$  linearized with respect to  $W(t)$ ,

$$\dot{W} = LW \quad (4)$$

where  $L$  may be a linear differential (or integral) linear operator.

The solutions are of the form

$$W(x, t) = e^{\lambda_k t} \sum_{d=0}^D t^d v_{k,d}(x) \quad (5)$$

where  $D > 0$  may happen if the corresponding eigenvalue  $\lambda_k$  is degenerate. In the following we consider  $D = 0$  and

$v_{k,d} = v_k$ . The unstable modes  $v_k \equiv v_u$  are connected with

$$\text{Re } \lambda_k \geq 0, \quad (6)$$

the stable modes  $v_k \equiv v_s$  with

$$\text{Re } \lambda_k < 0. \quad (7)$$

It is assumed that  $\text{Re } \lambda_k < A < 0$ ,  $A$  fixed, if the eigenvalues are discrete.

We decompose the wanted solution to the original non-linear and stochastic equations into a super position of modes determined by the instability analysis whereby we distinguish between the unstable and stable modes. The amplitudes of the unstable modes are the order parameters. Inserting

$$q(t) = q_0 + \sum_u \xi_u(t) v_u(x) + \sum_s \xi_s(t) v_s(x) \quad (8)$$

into the Eqs. (1) and projecting both sides of the resulting equation on the stable and unstable modes, we obtain equations of the form

$$\dot{\xi}_u = \lambda_u \xi_u + \hat{N}_u(\{\xi_u\}, \{\xi_s\}) + \hat{F}_u(\{\xi_u\}, \{\xi_s\}) \quad (9)$$

$$\dot{\xi}_s = \lambda_s \xi_s + \hat{N}_s(\{\xi_u\}, \{\xi_s\}) + \hat{F}_s(\{\xi_u\}, \{\xi_s\}). \quad (10)$$

$\lambda_u, \lambda_s$  are the eigenvalues (6), (7), which are assumed to be discrete. By a suitable, in general nonlinear, transformation to new variables,  $\hat{N}(\{\xi_u\})$  can be cast into a particularly simple form (“normal form” theory [101,103], initiated by Poincaré [113]).

If the eigenvalues  $\lambda_u, 0 > \text{Re } \lambda_s > -|B|$  are a continuous function of an index, e. g. a wave number  $k$ , wave packets of  $\xi_u(t)$  are used as new order parameter variables  $\mathcal{E}$  and  $\lambda_u(k)$  is replaced by an operator  $\Lambda_u(-i \frac{d}{dx})$  in one space-dimension or, more generally,  $\Lambda_u = (-i \nabla)$  [62]. For a related approach in fluid dynamics cf. [105].

The central idea of further procedure consists in eliminating the amplitudes of the stable modes. This is achieved by the *slaving principle* [56,62,65,148] which allows us to express the amplitudes of the stable modes in terms of the unstable modes

$$\xi_s(t) = f_s(\{\xi_u(t)\}, t), \quad (11)$$

where  $\xi_s, \xi_u$  are taken at the same time  $t$ . The explicit time-dependence of  $f_s$  stems exclusively from that of the fluctuating forces.  $f_s$  can be explicitly calculated in terms of a series expansion in powers of the order parameters. For practical purposes, in general only a few terms are needed.



For a general discussion of the convergence of this series see [62]. When noise is neglected, contact can be made with center manifold theory [76,112], which originally was a mere existence theory and was not constructive. For more recent developments, see books on bifurcation theory. A related approach is based on time-scale separation: The slowly damped or undamped modes serve as order parameters, which enslave the rapidly damped modes. A special case is adiabatic elimination.

**Resulting Langevin Equations** The enslaved mode amplitudes can be expressed by the order parameters and inserted in (9), so that closed equations for the order parameters alone result.

$$\dot{\xi}_u = \lambda_u \xi_u + \tilde{N}_u(\{\xi_u\}) + \tilde{F}_u(\{\xi_u\}, t) \quad (12)$$

where  $\tilde{N}$  is a polynomial of  $\xi(x, t)$  starting with at least second order.  $\tilde{F}$  is a stochastic force. A simple, yet prototypical example is (with a single order parameter  $\xi = \xi_u$ )

$$\dot{\xi} = \lambda \xi + a \xi^2 - b \xi^3 + F(t), \quad b > 0 \quad (13)$$

or

$$\dot{\xi} = -\frac{\partial V(\xi)}{\partial \xi} + F(t), \quad (14)$$

with the potential

$$V = -\frac{\lambda}{2} \xi^2 - \frac{a}{3} \xi^3 + \frac{b}{4} \xi^4. \quad (15)$$

If  $\lambda_u, \lambda_s$  (6), (7) represent a continuous spectrum, (generalized) Ginzburg–Landau equations result [62]. For example, the complex Ginzburg–Landau equation with fluctuating force reads [6].

$$\begin{aligned} (\xi(x, t) \equiv \xi_u, \text{ complex order parameter}) \\ \dot{\xi} = \lambda \xi + a \Delta \xi - c |\xi|^2 \xi + F(t). \end{aligned} \quad (16)$$

A further example is given by the Swift–Hohenberg equation [135], see also [24] (which was derived differently, however)

$$\dot{\xi}(x, t) = (a - b \Delta)^2 \xi(x, t) + c \xi(x, t) - d \xi(x, t)^3. \quad (17)$$

The Eqs. (12,13,16,17) allow for a great variety of solutions. In the case of real  $\lambda$  and a single order parameter, a nonequilibrium phase transition occurs (see below). In case of  $\lambda$  complex, and (at least) one complex order parameter, Landau–Hopf bifurcation [67,68], i. e. formation of a limit cycle may happen. In case of (at least) three order parameters and no noise, deterministic chaos may occur [89,118,132] (in the presence of noise, mixed effects may occur).

**Fokker–Planck Equation** Below and above the instability point in control parameters space, in a first step the fluctuations can be neglected and then, in the next step, taken care of by means of lowest order perturbation theory. In order to cover the transition region, under well defined conditions a Fokker–Planck equation for the probability density function  $f(\{\xi_u\})$  of the order parameters can be derived. For details see [62,63a] and the article by T. Frank, this volume.

The Fokker–Planck equation is of the general form

$$\dot{f}(\{\xi_u\}) = - \sum_u \frac{\partial}{\partial \xi_u} (\tilde{N}_u f) + \frac{1}{2} \sum_{uv} \frac{\partial^2}{\partial \xi_u \partial \xi_v} (Q_{uv} f). \quad (18)$$

It is assumed that  $\tilde{F}_u$  in (12) is  $\delta$  correlated in time,

$$\langle \tilde{F}_u(t) \tilde{F}_v(t') \rangle = Q_{uv} \delta(t - t'). \quad (19)$$

If  $\tilde{F}_u$  depends on  $\xi_u$ , the  $\hat{I}$  to, Stratonovich or Klimontovich procedure must be applied.

In the case of a single order parameter, where the Langevin equation [88], originally with  $\tilde{N} = -\alpha \xi$  is given by

$$\dot{\xi} = \tilde{N}(\xi) + F(t), \quad \langle F(t)F(t') \rangle = Q \delta(t - t'). \quad (20)$$

The steady state distribution function of (18) is given by [62,116]

$$f(\xi) = N \exp \left( -2 \int^{\xi} (\tilde{N}(\xi')/Q d\xi') \equiv N \exp(-2V(\xi)/Q) \right) \quad (21)$$

provided the boundary conditions are

$$f(\xi) \rightarrow 0 \quad \text{for } |\xi| \rightarrow \infty. \quad (22)$$

In the second Eq. (21),  $Q = \text{const.}$  is assumed.  $N$  is a normalization constant. A generalization of (18) to continuous variables,  $\xi_u(x, t)$ , gives rise to a functional Fokker–Planck equation. An explicit solution of the Fokker–Planck equation in the case of several discrete or continuous order parameters can be found if the drift and diffusion coefficients obey the rules of detailed balance [45,47].

**Nonequilibrium Phase Transition. Connection with Landau Theory** The explicit form of the solution of the Fokker–Planck Eq. (21) allows us to make contact with the theory of phase transitions in the sense of the Landau theory [86] where

$$\begin{aligned} f(\xi) &= N \exp(-F(\xi, T)/(kT)), \\ F(\xi, T) &= F(0, T) + a(T - T_c) \xi^2 + \frac{\beta}{4} \xi^4. \end{aligned} \quad (23)$$

In (21),  $V$  corresponds to the free energy  $F$  and the noise strength  $Q$  corresponds to *absolute* temperature  $T$ .  $T_c$  is the critical temperature, and (23) refers to a second order phase transition. In case of a first order phase transition, an additional term  $\gamma\xi^3$  appears in (23).

An important difference between phase transitions at thermal equilibrium and in the present case of non-equilibrium should be mentioned, however. The decisive constants in the case of non-equilibrium [62] phase transitions are rate constants in contrast to thermodynamic quantities in (23). While non-equilibrium phase transitions described by (21) were experimentally very well verified for instance in the case of lasers [116] (Fig. 1), in the case of thermal equilibrium the Landau theory can not be considered as a good approximation and had been replaced by the concept of critical exponents etc. as dealt with by renormalization group theory [75,146]. For a treatment of the time dependent Fokker–Planck equations see Risken [117].

In a number of cases the drift- and diffusion coefficients of the Fokker–Planck equation are by themselves expectation values, defined on the probability density function so that the Fokker–Planck equation becomes non-linear. For more details see the article by T.D. Frank in this volume.

**Instability of a Limit Cycle,  $q_0(t)$  [62]** The instability is checked by linear stability analysis by means of the hypothesis

$$q(t) = q_0(t) + W(t), \quad (24)$$

where  $q_0(t)$  is a time-periodic solution to (1) with  $\alpha = \alpha_0$ ,  $W(t)$  a small deviation.

Inserting (24) into (1) with  $F \equiv 0$  and linearization leads to an equation of the form (4), where  $L$  because of  $q_0(t)$  has become also a time-periodic function with the same period as  $q_0(t)$ . According to Floquet theory [35], the solutions to (4) with periodic  $L(t)$  are given by

$$W(t) = e^{\lambda_j t} v_j(t) \quad (25)$$

(in the case of nondegeneracy), where  $v_j(t)$  has the same period as  $q_0$ , i. e.  $L$ .

Depending on  $\text{Re } \lambda_j \geq 0$  or  $< 0$  we distinguish between unstable and stable modes (6, 7), respectively. One eigenvalue is  $= 0$  and corresponds to an indeterminate phase shift, which in nonlinear analysis is taken care of by a phase  $\phi(t)$  that acts as additional order parameter. In order to solve the fully nonlinear and stochastic equations,

the hypothesis

$$q(t) = q_0(t + \phi(t)) + \sum_u \xi_u(t) v_u(t + \phi(t)) + \sum_s \xi_s(t) v_s(t + \phi(t)) \quad (26)$$

is inserted in the Eqs. (1). The subsequent procedure follows the lines outlined above and leads to order parameter equations of the form

$$\dot{\xi}_u = \lambda_u \xi_u + \hat{N}_u(\{\xi_u\}, \phi) + \hat{F}_u(\{\xi_u\}, \phi) \quad (27)$$

$$\dot{\phi} = M(\{\xi_u\}, \phi) + G(\{\xi_u\}, \phi) \quad (28)$$

where  $\hat{N}, \hat{F}, M, G$  are polynomials in  $\{\xi_u\}$  and periodic functions of  $\phi$ .

The novelty as compared to the case of an unstable fixed point consists in the introduction of a phase as order parameter.

When noise is neglected, the newly evolving, i. e. bifurcating solutions are either two (or several) limit cycles or tori. Also basically, depending on the system, also a “back bifurcation” to a stable focus can happen.

**Instability of Tori [62]** The corresponding theory is rather complex so that a few words must suffice here. The basic idea [62] is based on an extension of (24,26) where  $q_0$  is chosen as a quasi periodic function

$$q_0 = q_0(\omega_1 t, \omega_2 t, \dots, \omega_M t) \quad (29)$$

where the  $\omega$ 's must be sufficiently irrational in the sense of the KAM (Kolmogorov [80], Arnold [3], Moser [100]) theorem. Besides amplitudes as order parameters, also phases  $\phi_1(t), \dots, \phi_M(t)$  are introduced. For details cf. [62], and for alternative approaches [21,128].

### A Remark on the Method of Solution of Evolution Eqs. (1)

In this article the central role of order parameters is stressed because this allows us to establish profound analogies between quite different systems. In practical applications it may be preferable, however, to apply other methods of solution, analytical, numerical or mixed, in order to derive the spatial, temporal or spatio-temporal patterns. In this way, the Springer Series in Synergetics have developed a “tool box” of models [98].

### Quantum Theoretical Formulation

In a quantum theoretical treatment one deals with quantum mechanical Langevin equations which are Heisenberg equations of motion for operators to which pumping

and damping terms as well as random noise sources are added. Here, according to quantum theory, the system's observables are represented by time-dependent quantum mechanical operators,  $\Omega_j$ . For instance, by the position operator  $\hat{x}$  and the momentum operator  $\hat{p}$  of a particle, or, in quantum field theory, by creation and annihilation operators  $\hat{b}^+$ ,  $\hat{b}$ , respectively. The quantum mechanical Langevin equations read (see, for instance [54,58]):

$$\dot{\Omega}_j = \frac{i}{\hbar} [H, \Omega_j] + \text{damping} + F_j(t), \quad (30)$$

where  $H$  is the Hamilton operator, and  $F_j(t)$  are stochastic operators which usually are assumed to be  $\delta$ -correlated in time. The quantum mechanical properties can be determined by the postulate of quantum mechanical consistency of  $\Omega_j$ , (cf. [54], appendix).

If the non-commutativity of operators is taken care of, the procedure to derive order parameter equations is formally the same as in the case of classical Langevin equations as indicated above. The Fokker-Planck equation, however, must be replaced by a density matrix equation, originally introduced as master equation [109]. For nonequilibrium systems, such as the laser, see [50,126,144], also [54,119]. Using methods of quantum classical correspondence, this density matrix equation can be converted into a Fokker-Planck equation under specific conditions. The basic idea is this:

### Quantum-Classical Correspondence

There are several ways to define quantum classical correspondence. In the case of position operator  $\hat{x}$  and momentum operator  $\hat{p}$  with the commutator  $[\hat{p}, \hat{x}] = \frac{\hbar}{i}$  and the density matrix  $\rho$ , the Wigner distribution function  $W(x, p)$  [145] is defined by

$$W(x, p) = \frac{1}{(2\pi)^2} \cdot \int \int_{-\infty}^{\infty} e^{-ikx-ilp} \cdot \text{tr} \left( e^{ik\hat{x}+il\hat{p}} \rho \right) dk dl \quad (31)$$

where “tr” means trace.

Thus a relation is established between the quantum mechanical density matrix and a classical quasi-density  $W(x, p)$ . Based on (31) or (34,35,36), a density matrix equation can be converted into a generalized Fokker-Planck equation [54].

By the transformation of  $\hat{x}$ ,  $\hat{p}$  to creation and annihilation operators  $b^+$ ,  $b$  by means of

$$\hat{b}^+ = \frac{1}{\sqrt{2\hbar}} (\hat{x} + i\hat{p}) \quad (32)$$

$$\hat{b} = \frac{1}{\sqrt{2\hbar}} (\hat{x} - i\hat{p}) \quad (33)$$

an alternative form to (31) is given by

$$P(\beta, \beta^*) = \frac{1}{\pi^2} \int \int_{-\infty}^{\infty} e^{-i\beta k - i\beta^* l} \cdot \text{tr} (e^{ik\hat{b}^+ + il\hat{b}} \rho) dk dl. \quad (34)$$

Because  $\hat{b}^+$ ,  $\hat{b}$  are noncommuting operators,  $[\hat{b}^+, \hat{b}] = 1$ , different “quasiprobability” distributions  $P$  result, if

$$e^{ik\hat{b}^+ + il\hat{b}} \quad (35)$$

is replaced by

$$e^{ik\hat{b}^+} e^{il\hat{b}} \quad (35)$$

or

$$e^{ik\hat{b}} e^{il\hat{b}^+}. \quad (36)$$

(35) gives rise to the Glauber-Sudarshan representation. For details and references see [54].

### Regular Spatial and Spatio-Temporal Patterns

One of the most striking features of nonequilibrium systems in physics, chemistry and biology is their capability of forming (more or less) regular spatial pattern (for explicit examples see below). (There is a rich literature on pattern formation in physics, especially fluids [20,24,92,136], but also semiconductors [125] and nonlinear optics [133], chemistry [28,33,83] and biology [7,95,96,102] and general [69,70,98,104,110,111,115,140]. Furthermore, the patterns exhibit striking similarities in spite of the fact that the individual parts are quite different. The methodology of Synergetics (e.g. [62]) provides us with a basic insight into the causes of such analogies.

Pattern formation is determined by at least three causes:

1. internal mechanism, such as e.g. the interplay between reactions and diffusion in large scale chemical processes,
2. the influence of boundaries,
3. initial conditions.

Concerning 1) and 2) between two (limiting) cases can be distinguished.

1. dimensions of the internally evolving patterns are of the same or larger order as those of the boundaries. Here a strong influence of the boundaries must be expected.

2. dimensions of evolving patterns are small compared to those of the boundaries (boundaries  $\rightarrow \infty$ ).

To bring out the essential features we consider that originally for a control parameter value  $\alpha_0$  the system is homogeneous and quiescent. The approach can, however, be extended to a space dependent reference state (which, e. g. resulted from a first bifurcation leading to  $q_0 = q_0(x)$ ) and the cases of a limit cycle or torus. The space may be 1, 2 or three dimensional Euclidian or, e. g., a 2 or 3 sphere.

### Infinite Boundaries

We start with 2) infinite boundaries, the medium is homogeneous and isotropic. We assume a continuous transition from the homogeneous to the “bifurcating” state. The evolving patterns are determined by the leading terms in (8) that we call the “mode skeleton”

$$q(x, t) = q_0 + \sum \xi_u(t) v_u(x) \quad (37)$$

and the order parameter Eq. (12). The functions  $v_u(x)$  are the *space-dependent part of the solutions* to (4) where  $L$  is a differential (or integral) operator which is invariant against translation and rotation. Thus, e. g.,  $L$  commutes with the displacement operator

$$\Omega_a: x \rightarrow x + a, \quad a \text{ constant vector.}$$

Thus  $v_u(x)$  can be chosen as eigenfunction to  $\Omega_a$ ,

$$\Omega_a v_u(x) = \Lambda v_u(x) \quad (38)$$

with

$$v_u = e^{ikx} \quad (39)$$

$$\Lambda = e^{ika} \quad (40)$$

i. e. plane waves. Which waves must be considered in (37) is determined by  $\lambda_u$  in (6) as well as by the order parameter Eq. (12).

The condition  $\text{Re } \lambda_u(k) = 0$  defines  $k = k_{\text{crit}}$ . As was shown by means of many examples  $k \approx k_{\text{crit}} \neq 0$ . If the boundaries are finite, such a discrete  $k$  must be chosen which comes closest to  $k_{\text{crit}}$ . If the boundaries tend to infinity, a continuous set  $k$  is taken care of by (generalized) Ginzburg–Landau equation (see above). If the boundaries are “narrow” in 1 or 2 dimensions, but large in the remaining dimensions, the wave vector  $k$  must be split into  $k_{\parallel}$  and  $k_{\perp}$  where  $k_{\parallel}$  is practically continuous and  $k_{\perp}$  discrete. Quite often only one  $k_{\perp}$  (the most critical) needs to be considered. This leads to practically 2 (or 1) dimensional

patterns connected with  $k_{\parallel}$ . In the 2-dimensional case, the modes with  $|k_{\parallel}| = k_{\text{crit}}$  are degenerate. This degeneracy can be lifted by a weak influence of boundaries (leading to roll patterns), by specific initial condition which (by chance) prefers a specific roll pattern, or by terms in the order parameter-equations that lead to specific combinations, e. g.

$$k_1 + k_2 + k_3 = 0, \quad (41)$$

where  $k_j, j = 1, 2, 3$  belong to  $k_{\parallel}$ .

This gives rise to the formation of hexagons. This is the case if the leading term of  $\tilde{N}$  contains

$$\int v_{k_1} v_{k_2} v_{k_3} d^2x \neq 0. \quad (42)$$

In three dimensions this mechanism may lead to plane wave fronts stabilizing each other which gives rise to icosaeders, as observed in diatomea.

An important class of spatio-temporal patterns (in 2 dimensions) results when the system utilizes *rotation symmetry*. This can best be explained by the following example:

In many cases of practical interest,  $N$  in (1) and thus  $L$  in (4) contain the Laplace operator  $\Delta$ . When written in planar polar coordinates  $r, \vartheta$ , solutions to (4) are of the general form

$$v \propto e^{i(m\vartheta - kr - \omega t)} \quad (43)$$

(times a rotation symmetric function  $g(r)$ ) which represents *spirals*.  $m = 0$  represents concentric rings, while an integer  $m > 0$  represents the number of spiral arms.  $\omega = 0$  represents standing spirals,  $\omega \neq 0$  rotating spirals.

The mode skeleton (37) is composed of functions of the form (43). Which of the functions (43) appear in (37) depends on the competition Eqs. (12) for order parameters, which may also allow for a super position of counter rotating spirals (such as in the sunflower head). As group theory shows (see below), solutions (43) with different  $m$ 's belong to different irreducible representations, and do not coexist in (37). This does not exclude the coexistence of differently rotating spirals in *different* regions of space, however.

The above results can be cast into the isomorphy principle:

While the “true”  $q$  is represented by (we omit the homogeneous  $q_0$ )

$$q = \sum_u \xi_u v_u(x) + \text{enslaved modes, with same symmetry.} \quad (44)$$

and  $v_u$  “true modes”, its symmetry features can be replaced by a “representative”  $q'$ :

$$q' = \sum_k \xi_k R_k(x), \quad (45)$$

where  $R_k$  represent the “elementary” functions showing the symmetry under consideration. While the material significance and explicit form of  $q$  according to (44) may be quite different for different material substrates,  $q'$  (45) shows the *same* patterns for different systems.

These results can be deepened by invoking group theory, in which also the effect of the boundaries is taken into account.

### Symmetries, Group Theory, Representation Theory, Finite Boundaries

Consider a set of transformations  $G_j$  of space variables  $x \rightarrow x'$  so that

$$G_j q \rightarrow q' \quad (46)$$

*Example 1*  $G_j$  induces the translation

$$x \rightarrow x + a \quad \text{so that} \quad G_j q(x) = q(x + a). \quad (47)$$

The transformations must be so that they are compatible with the internal properties of the system (1) and the boundary conditions. Example: when dealing with a problem on a 2-dimensional sphere, the transformed coordinates  $x$  must not leave the sphere.

Because of the symmetry of the problem, the transformations  $G_j$  form a group defined by

1. existence of unity  $E$  such

$$G_j E = G_j \quad \text{for all } j \quad (48)$$

2. the product of two group elements is again an element of the group,

$$G_j G_k = G_l \quad \text{for all } j, k \quad (49)$$

3. existence of an inverse  $G_j^{-1}$  for all  $j$  so that

$$G_j^{-1} G_j = E, \quad (50)$$

4. associative law

$$(G_k G_l) G_j = G_k (G_l G_j) \quad (51)$$

for all group elements.

In the following we first ignore random forces, i. e. we consider (1) with  $F \equiv 0$ .

$$\dot{q}(x, t) = N(q, \Delta, \alpha). \quad (52)$$

Jointly with the boundary conditions, (52) defines a function space  $S$  in which all functions to be considered must lie (i. e. can be represented by linear combinations of a complete set of (vector valued) basic functions of  $S$ ; example:  $S$  is a Hilbert space)

**Definition 1** The system is invariant against  $G_j$  if for all  $f \in S$

$$G_j N(G_j^{-1} f) = N(f). \quad (53)$$

*Example 2*

$$G_j: x \rightarrow x + a, \quad (54)$$

$$N(f) = \Delta f + V(x)f + f^2. \quad (55)$$

Then

$$G_j \cdot N(G_j^{-1} f) = \Delta G_j^{-1} f(x+a) + V(x+a) G_j^{-1} f(x+a) + (G_j^{-1} f(x+a))^2 \quad (56)$$

$$= \Delta f(x) + V(x+a)f(x) + f(x)^2 \quad (57)$$

$$\neq N(f) = \Delta f(x) + V(x)f(x) + f(x)^2 \quad (58)$$

unless  $V(x+a) = V(x)$ . If  $a$  in (54) is arbitrary,  $N$  is not invariant against (54).

Application of  $G_j$  to  $q$  in (52) leads to

$$\frac{d}{dt}(G_j q) = N(G_j q) \quad (59)$$

or because of (53), (with  $f = G_j q$ ), to

$$\frac{d}{dt}(G_j q) = G_j N(q). \quad (60)$$

In the spirit of representation theory of groups the action of  $G_j$  on  $f$  can be understood as an abstract operation, but also as a matrix acting on the vector  $f$  in  $S$ -space.

By appropriate transformation of basis of  $q$ , and using the representation theory of symmetry groups, all matrices  $U_j$  belonging to all group elements  $j$  can simultaneously be decomposed into “irreducible” representations so that (in the example of 3 irreducible representations)

$$U_j = \begin{pmatrix} \square & \bigcirc & \bigcirc \\ \bigcirc & \square & \bigcirc \\ \bigcirc & \bigcirc & \square \end{pmatrix} \quad (61)$$



Each box  $\square$  is a matrix  $U_j^{(k)}$  with dimension  $Dk$ , so that

$$D1 + D2 + \dots + Dk = \text{dimension } U_j.$$

**Example 3** Rotation group applied to 2-sphere (e. g. earth surface). Basis functions are spherical harmonics  $Y_m^l$  with “quantum numbers”  $l, m$ . Subspace  $l$  fixed,  $m = 0, \dots, l-1$ . As a consequence, the mode skeleton reduces to ( $q_0$  dropped)

$$q^l = \sum_m \xi_m(t) Y_m^l. \quad (62)$$

There is no coupling between different  $l$ s, which implies a low dimensional dynamics of  $\xi_m$ .

Generally, the original function space  $S$  is decomposed into subspaces forming the basis of each irreducible representation. This implies a symmetry reduction beyond bifurcation point, compared to the situation below bifurcation point, where

$$G_j q = q \quad \text{for all } j, \quad (63)$$

i. e.  $q$  fully symmetric under  $G$ .

In our example beyond the bifurcation point  $q$  is given by  $q^l$  where  $Y_m^l$  transforms according to the subgroup  $G^l$ , which leaves the space spanned by  $Y_m^l$  invariant. If, however, group elements not belonging to  $G^l$  are applied to  $q^l$ , this space is left. In other words,  $q^l$  is connected with a lower symmetry than  $q$  (63). By bifurcations, the symmetry of  $q$  is lowered and one speaks of “symmetry breaking instability”. If fluctuating forces in (1), i. e. in (52) are taken into account, the full symmetry can be restored (under specific conditions on the fluctuating forces).

While group theory has found important and widespread applications to quantum theory, it is less frequently used in problems of Synergetics, though there it may lead to deep insights as pointed out above. (For an in-depth approach see [42,43,120].)

On top of, or jointly with, regular patterns, a variety of defects as well as boundaries between different patterns may occur (cf. contribution by Pismen, this volume and [110,111]).

### A Further Mathematical Tool: Shannon Information and the Maximum (Information) Entropy Principle

While evolution equations are the backbone of Synergetics, also other tools are invoked to deal with complex systems. Such a tool is Shannon information [129] which is defined by

$$i = - \sum_j p_j \log_2 p_j \quad (64)$$

where  $p_j$  is the relative frequency of the event  $j$  or, in a different interpretation, the probability of finding the realization  $j$  in an experiment. The maximum (information) entropy principle as formulated by Jaynes [73,74], for an earlier proposal see [27]), allows one to make unbiased guesses on systems on which only incomplete data are known by maximizing the informations, i. e. (64) = max! or = extremum! under given constraints.

A simple example is provided by a gas composed of  $N$  particles, where the total kinetic energy  $E_{\text{kin}}^{\text{tot}}$  is fixed. Denoting the kinetic energy of a particle with mass  $m$  and velocity  $v_i$  by  $f_i = (m/2)v_i^2$ , the mean kinetic energy per particle is

$$\sum_i p_i f_i = E_{\text{kin}}^{\text{tot}}/N \quad (65)$$

To fix  $p_i$ , (64) must be maximized under the normalization condition

$$\sum_i p_i = 1 \quad (66)$$

and the constraint (65).

Using Lagrange multipliers,  $\lambda, \lambda_1$ , the result reads

$$p_i = \exp(-\lambda - \lambda_1 m v_i^2/2) \quad (67)$$

i. e. the Maxwell-Boltzmann distribution function. Also relations between the Lagrange multipliers  $\lambda, \lambda_1$  can be established which, evidently, have fundamental thermodynamic significance.

This approach has been extended to the treatment of nonequilibrium phase transitions, i. e. determination of order parameters, enslaved modes and emerging patterns [60]. The crucial idea consists in the proper choice of constraints, as which the moments of the variables  $q_i$  are chosen:

$\langle \dots \rangle$  means average over the joint distribution function  $f(q_1, q_2, \dots, q_n)$  which replaces  $p_j$  and the vector  $(q_1, \dots, q_N)$  replaces  $j$ . The variables  $q_j$  may be discrete or continuous.

$$f_i = \langle q_i \rangle, \quad i = 1, 2, \dots, N. \quad (68)$$

$$f_{ij} = \langle q_i q_j \rangle. \quad (69)$$

$$f_{ijkl} = \langle q_i q_j q_k q_l \rangle, \quad i, j, k, l = 1, 2, \dots, N. \quad (70)$$

The resulting distribution function is given by

$$q = \exp V(\lambda, q) \quad (71)$$

with

$$V(\lambda, q) = \lambda + \sum_i \lambda_i q_i + \dots + \sum_{ijkl} \lambda_{ijkl} q_i q_j q_k q_l. \quad (72)$$

(71) is a starting point to make contact with the Landau or Ginzburg–Landau theory of phase transitions [86], and to guessing Fokker–Planck equations. The approach allows one to calculate the efficiency of self-organizing systems close to their instability points.

The method has been extended to the “unbiased modeling” of stochastic processes: how to guess path integrals, Fokker–Planck equations and Langevin–Itô equations [60]. The central quantity to be searched for is the probability density  $P_n$  of paths.

Let  $q(t)$  be the state vector  $q = (q_1, \dots, q_N)$  at time  $t$ , then

$$P_n(t_n, t_{n-1}, \dots, t_0) = P_n(q(t_n), t_n; q(t_{n-1}), t_{n-1}; \dots; q(t_0), t_0), \quad t_n > t_{n-1} > \dots > t_0. \quad (73)$$

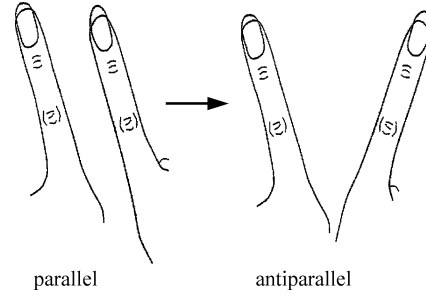
This task is simplified if the Markov hypothesis on the process holds, i. e.

$$P_n(t_n, t_{n-1}, \dots, t_0) = \hat{P}(q(t_n) | q(t_{n-1}) t_{n-1}) \cdot P_{n-1} \quad (74)$$

where  $\hat{P}$  is the transition probability so that only transition probabilities between subsequent states (with  $\Delta t \rightarrow 0$ ) must be guessed in addition to  $P_0$ . In the frame of the present approach, this task is fulfilled by use of the maximum information principle. The constraints to be used are essentially conditional first order moments and two-time correlation functions of the state vectors  $q(t)$ ,  $q(t')$ .

### Phenomenological Synergetics

In many fields of science, including medicine, the microscopic variables and their dynamics are not well-known or not known at all. Nevertheless, in quite a number of cases, namely where dramatic macroscopic changes of the system’s behavior take place, general insights, gained by Synergetics, can be invoked. A paradigm for this procedure is the modeling of Kelso’s finger experiments [77,78] (Fig. 3). He instructed subjects to move their index fingers in parallel which was accordingly performed. However, when the speed of the fingers was increased, the parallel movement was replaced by a symmetric movement quite involuntarily and spontaneously. In other words, a transition from a parallel to an anti-parallel phase takes place. In terms of Synergetics, the interpretation is simple: the control parameter consists in the prescribed frequency  $\omega$  of the finger movement, whereas the macroscopic quantity, i. e. the



**Synergetics: Basic Concepts, Figure 3**

**Transition between finger movements from parallel to symmetric in Kelso’s experiment [64]**

order parameter that changes dramatically is provided by the relative phase of the two index fingers. According to the experience made in Synergetics, the order parameter, here called  $\phi$  obeys a typical order parameter equations of the form [64]

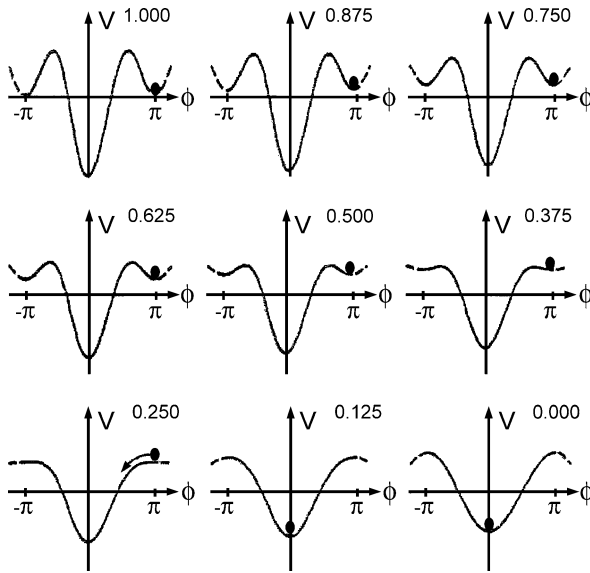
$$\dot{\phi} = -\frac{\partial V}{\partial \phi} + F(t), \quad (75)$$

where  $V(\phi, \omega)$  is a potential function and  $F$  a fluctuating force. When the control parameter  $\omega$  is changed, the potential runs through a series of forms as depicted in Fig. 4. As was shown in detail, at a critical value of  $\omega$ , the transition from one potential minimum to another one occurs, as related to the change of the kind of finger movement. The mathematical analysis shows hysteresis, critical slowing down and critical fluctuations [59] which reject the idea that the brain acts like a computer via a motor program but rather via self-organization.

Another application is made by the Synergetic computer [61] (Figs. 5, 6), where to each pattern to be recognized a specific order parameter is attached. Pattern recognition is then achieved via a competition between order parameters. The competition equations are given by

$$\dot{\xi}_k = \frac{\partial V}{\partial \xi_k} V(\xi_1, \dots, \xi_M) = -\frac{1}{2} \sum_k \lambda_k \xi_k^2 + \beta \sum_{k,k'} \xi_k^2 \xi_{k'}^2 - C \sum_k \xi_k^4. \quad (76)$$

This approach may serve also for modeling of brain functions: both recognition as well as movements are governed by the establishing of order parameters which may wander from one quasi attractor to another one. Quasi attractors are defined as attractors that vanish after the task has been accomplished, e. g. after a pattern has been recognized or movement performed.



**Synergetics: Basic Concepts, Figure 4**

Sequence of potential curves of the Haken–Kelso–Bunz model of Kelso's experiment [64]

Based on the concept of order parameters, a learning procedure for Synergetic computers has been developed [61]. Here the number of patterns to be recognized is prescribed and then a special functional must be min-

imized. In the case of the Synergetic computer, it is possible to make contact between the microscopic and the mesoscopic description, i. e. the microscopic variables are pixel values  $q_j$ ,  $j$  pixel index, whereas the mesoscopic (or macroscopic) quantities are the order parameters  $\xi_k$ .

The relation between  $\xi_k$ ,  $q_j$  is given by

$$\xi_k = \sum_j v_j^{k+} q_j, \quad (77)$$

where  $v_j^{k+}$  are adjoint prototype patterns, with  $k$  pattern index,  $j$  pixel index.

The relation between prototype patterns  $v_j^k$  and their adjoints is given by

$$\sum_j v_j^{k+} v_j^{k'} = \delta_{kk'} \quad (78)$$

At the phenomenological level the order parameter concept allows us to interpret and model complex movement patterns, e. g. learning to ride on a pedalo [59]. In the experiments, LED's are fixed at the joints of the subject and their positions measured which gives rise to a series of time-dependent tracks. Then, in a first step, a principle component analysis is performed, in the next step, by means of a variational principle, the best fit is searched in



**Synergetics: Basic Concepts, Figure 5**

Recognition of faces by the synergetic computer: stored or learned prototype patterns [61]



**Synergetics: Basic Concepts, Figure 6**

Pattern recognition by the synergetic computer: recognition of a specific face of which initially only a subset of pixels is presented [61]

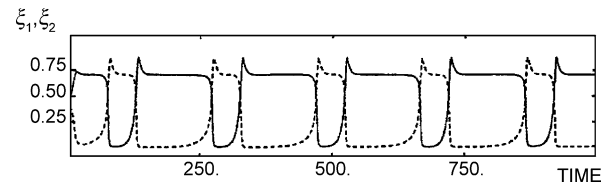


**Synergetics: Basic Concepts, Figure 7**

Example of an ambivalent figure: young / or old woman? [34]

terms of order parameters and their equations of motion, in order to mimic the actual tracks. While in the learning phase several order parameters are needed, at the end the whole movement is governed by a rather simple equation for a complex order parameter.

During the development of Synergetics it turned out that there are strong relations to gestalt theory [85] as well as to psycho physics. A typical example is provided by ambivalent figures where (Fig. 7) [34] shows an example. An observer may either perceive a young woman or an old woman, but not both simultaneously, rather the perception switches between these two percepts. In the mathematical modeling to each percept an order parameter is attached [61], which obeys the typical equations of Synergetics. The control parameter invoked here is attention. According to an early suggestion by Wolfgang Köhler [85], when a pattern is recognized, the corresponding attention fades away. This has been modeled mathematically based on a competition dynamics between two order parameters, when the control parameter (attention) of one pattern fades away, the other pattern gets the possibility of being perceived. Then in the next step the corresponding attention parameter fades away and the first pattern may re-appear (Fig. 8) [61]. This model describes details of the observed phenomena, such as the dependence of the duration of the perception of one face as compared to that of the other face, dependent on the bias which face is recognized first. Also, one may distinguish between slow, medium and fast observers, depending on the individual parameters.



**Synergetics: Basic Concepts, Figure 8**

Order parameter oscillations belonging to the recognition young woman / old woman with bias towards the young woman [59]

Quite generally, order parameters may have properties of gestalt in the sense that they are invariant against size, orientation and perception of objects in space.

In medicine, a syndrome has the characteristic features of an order parameter. On the one hand it is generated by the co-operation, or at least by the simultaneous presence of specific features, on the other hand once the syndrome (order parameter) is established, it acts on the individual parts of the system, where the slaving principle induces specific phenomena at the level of individual parts. Clearly, the concept of circular causality plays an important role here. It shows that the syndrome, at least in general, can not be cured by curing an individual symptom, but rather by curing a decisive majority of individual causes.

### Semantic Synergetics

In soft sciences, but also in medicine and other fields, a mathematical modeling, even at the level of order parameters may not be possible. Nevertheless, Synergetics may provide us with qualitative insights into basic mechanisms. In psychology and psychiatry [122], quite often specific mental states can be ascribed to a patient. For instance in bipolar patients a depressive phase or a manic phase may appear or in depressive patients a normal phase and a depressive phase. Another example is provided by patients with a compulsory action. In the spirit of Synergetics, as a theory of indirect control, one may ask, whether there are appropriate control parameters by means of which the behavior of a person can be changed. Let the two states be represented by the positions of a ball in a landscape with two valleys. In this situation, direct control means to push the ball from the unwanted position to the wanted. Indirect control means to lower the potential hill between the two valleys so that the wanted transition may occur via self-organization. This may happen through interventions used in cognitive psychology, a change of environmental conditions, or/and by specific medication. The central issue here is that the patient is not directly influenced, e. g. by saying you must do this or that, but rather by a soft

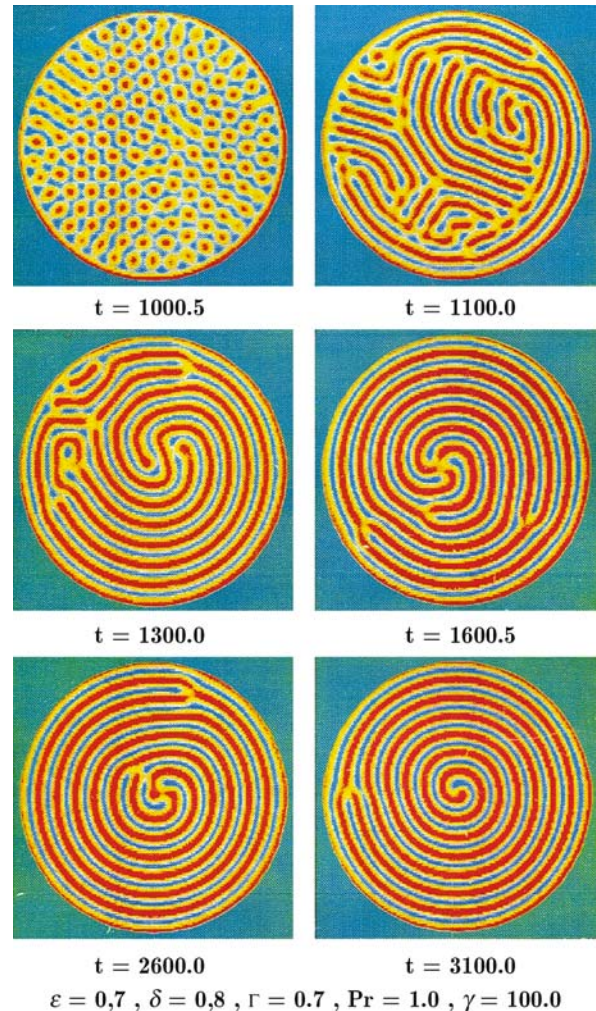


changing of his/her point of view. A number of successes have been reported about this method which is, to some extent, well known in psychiatry, but finds here a scientific theoretical basis. For more details see the article by G. Schiepek and V. Perltz, this section, and in a somewhat related form [66].

### Some Selected Examples

The study of nonlinear, self-sustained oscillations [1,2,13] be it in radio-engineering, mechanics or other fields, has a long tradition. In the context of bifurcation theory, their origin was unearthed by Hopf [67,68].

Nonlinear optics [99] and, when quantum effects are important, quantum optics [57,97,123,142] provide us with a wealth of phenomena, in particular of the formation of coherent oscillations. A device, closely related to the laser, is the parametric oscillator [44], in which, within a nonlinear crystal, incoming pump light is split into a signal and an idler. Then, similar to the laser light, the signal light becomes amplified, and its generation can be described as that of a nonlinear quantum-mechanical oscillator. Fluid dynamics is rich of pattern formations (including chaos) [12,16,17,18,31,41,89,92,105,115,118,127,135,136], to mention just a few. In a fluid heated uniformly from below, with increasing temperature difference, several instabilities may occur for instance giving rise to stationary patterns, such as rolls, hexagons (Fig. 9) or squares. In the next step the rolls may start to show oscillations, and still more complex patterns may occur (Fig. 3). In the case of the Taylor instability [137], a liquid is placed in between two coaxial cylinders, where the outer one is rotating. With increasing rotation speed, a hierarchy of instabilities is reached, first the formation of rolls, then oscillating rolls at one frequency, then oscillation of rolls at two frequencies, and finally weak turbulence, i. e. chaos occurs (Fig. 10) [31,93]. Important phenomena are the establishing of boundaries and of defects as described in the article by Pismen [110,111] and other articles of this Encyclopedia. A rich variety of pattern formation may occur in semiconductors [125], where electrons and holes as well as currents form specific spatio-temporal patterns. In meteorology, atmospheric convection patterns and other instabilities are treated [38]. In chemistry, oscillations and large scale patterns arise by means of the interplay of chemical reactions and diffusions [9,15,28,32,33,149], e. g. concentric ring patterns, each starting from a center, which then annihilate each other when colliding. An important class is provided by spiral patterns which may have one to several arms (Fig. 11). In biology, specific models on morphogenesis were treated, such as the formation of stripe or spot

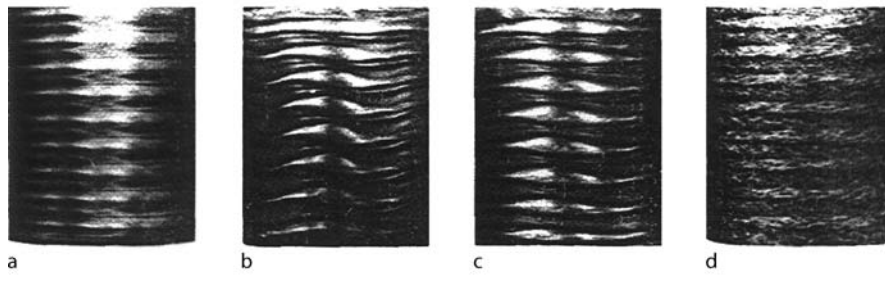


**Synergetics: Basic Concepts, Figure 9**

Model calculation of the motion of a fluid in a circular pan uniformly heated from below (after [29]). *Upper left corner:* above a critical temperature difference between lower and upper surface of the fluid layer, a hexagonal pattern appears. If the boundary is also heated uniformly, a transition to the spiral pattern with one or several arms can be found (*lower right corner*)

patterns on animal furs or skins of fish (Fig. 12) or still more complicated patterns on sea shells [39,62,95,96,102]. The basic idea which can be traced back to Turing [139] is this: originally unspecialized cells produce activator and inhibitor molecules which by reaction and diffusion form a prepattern, a morphogenetic field [147]. At positions of high activator concentration, genes are switched on which then leads to cell differentiation producing e. g. pigments. In aggregating slime mold, spiral or concentric ring patterns are observed [14,37]. Mathematical models on prebiotic evolution [26] study the competition between species of





**Synergetics: Basic Concepts, Figure 10**

Pattern hierarchy in the Taylor–Couette instability. A fluid in between two vertical coaxial cylinders of which the outer one rotates, shows no macroscopic movement pattern, if the movement of the outer cylinder is slow. When the rotation speed is increased, first a role pattern appears in which the fluid moves outwards at one height, and then inwards at another height. This movement pattern is periodic with respect to height [137]. At a further critical rotation speed, the pattern shows oscillations which at a further speed transform into a motion with two frequencies until eventually chaotic motion appears. The experiments were done by [31], the modeling was done for the first transition (homogeneous to roles) and especially the second transition (roles to oscillating roles) by [93]



**Synergetics: Basic Concepts, Figure 11**

Belousov–Zhabotinsky reaction: the occurrence of spirals (courtesy A.T. Winfree). They may show one to several arms. The centers of the spirals may occur at different positions. Spirals hitting each other, annihilate each other

biomolecules and the “survival” of the fittest, where pronounced analogies with the dynamics of laser photons can be unearthed, fully in line with Synergetics [62]. In the understanding of brain function, for instance, steering of movements, pattern recognition or decision making, the reduction of degrees of freedom of the numerous neurons to few order parameters is central [59].

The concepts and principles of Synergetics shed new light on important relationships in economy, such as co-operation and competition between companies, the im-



**Synergetics: Basic Concepts, Figure 12**

Stripe pattern on a tropical fish

portant role of indirect steering by means of control parameters, such as taxes, interest rates. It can be shown, that a fusion of companies does not necessarily lead to so called synergy effects, but rather critically depends on initial conditions and details of the cooperation between the previously separated firms. Important insights are also gained into fundamental processes of climatology, as well as in ecology such as the by now well-known and publicly discussed effects that even small concentrations of chemicals in the atmosphere can change the climate dramatically. The same is true for lakes, in which beyond a critical pollution, fish population dies out entirely.

In this way, the numerous examples collected in the field of Synergetics, provide not only scientists but also the public with impressive examples of dramatic changes (in-

stabilities) provoked by even a slight change of control parameters. Clearly, an important research subject of Synergetics is a detailed study of which control parameters are critical and to which control parameters a system is rather insensitive. Sociology is an important field for the application of stochastic models [8]. In particular, basic concepts of Synergetics have proven useful in the developing field of sociodynamics, where e. g. phase transition-like phenomena may occur [143].

### History and Relations to Other Fields

The term Synergetics was coined by H. Haken in 1969 in a lecture at University of Stuttgart. A first description of the goals of this field was given by H. Haken and R. Graham in 1971 [63] where the unifying role of the concept of order parameters is outlined. A relationship exists to the general system theory due to von Bertalanffy [10], which also aims at the exploration of analogies between different systems, but on the level of the individual elements rather than on the level of order parameters. Von Bertalanffy coined the term flux equilibrium (Fließgleichgewicht) in order to characterize homeostasis in active systems [11]. A general mathematical frame for Synergetics is provided by dynamic systems theory (see, for instance, [49]) which, however, in the traditional approach ignores stochastic processes (mainly chance events) which are also of great relevance for Synergetics. Here the theory of Markov processes with their typical equations, such as Langevin equations, Fokker–Planck equations, Chapman–Kolmogorov equations, the Kramers–Moyal expansion etc. is important (see for instance [134] and ► [Linear and Non-linear Fokker–Planck Equations](#) by T. Frank).

A basic feature of Synergetics consists in dealing with nonlinearities in complex systems and studying, mainly quantitatively, qualitative changes at macroscopic scales. Qualitative changes of systems at macroscopic levels are studied also by catastrophe theory [5,138], which may be interpreted as a study of the surfaces of equilibrium points of few order parameters, where different cases are classified according to the (low) number of control and order parameters. Chaos theory studies the mostly irregular dynamics of deterministic low dimensional continuous [89,106,118,132] or discrete dynamic systems [23,30,48,94,130], where the behavior is mainly characterized by so called Lyapounov exponents, various kinds of fractal dimensions and chaotic attractors. The slaving principle of Synergetics provides a basis for an application of chaos theory to multi-component systems in that Synergetics shows the possibility of reducing the degrees of freedom. Synergetics shares some of its topics with singularity the-

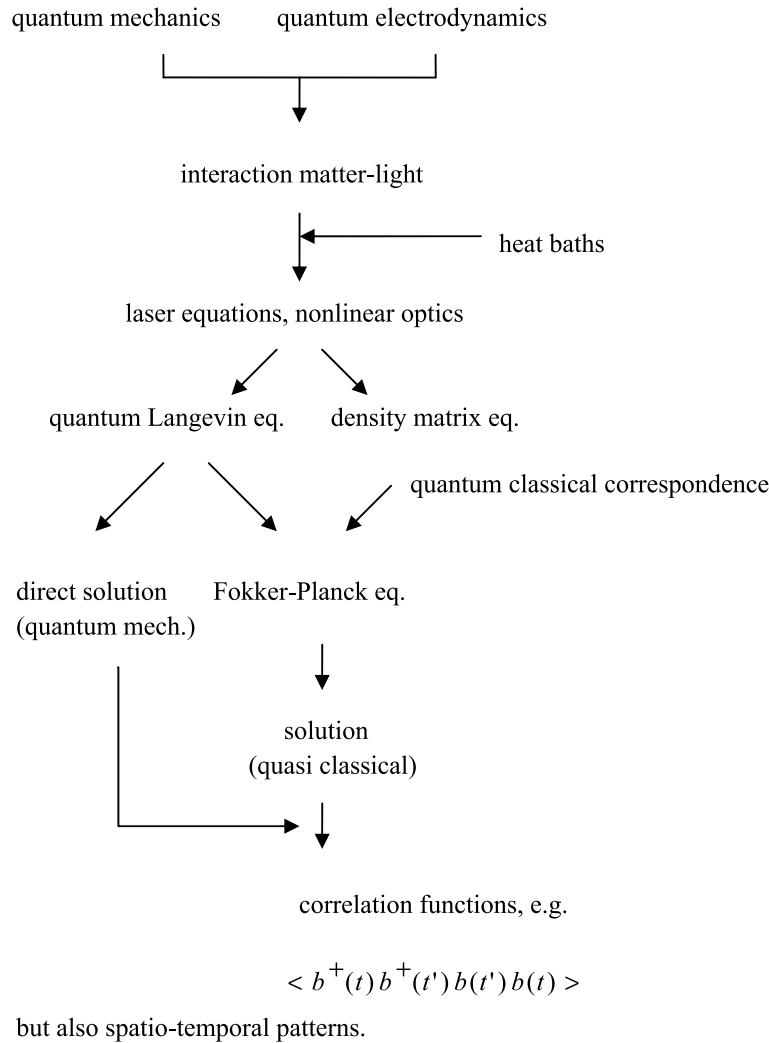
ory [4,42,43], which applies to bifurcation points and their surrounding. Another point of contact is bifurcation theory (see the quotations in previous chapters), in which the branching of solutions of the dynamic system close to instability points is studied. The term dissipative structure was coined by Prigogine [40] to characterize evolving structures in systems away from thermal equilibrium where as in all such non-equilibrium systems dissipation occurs. A typical example is that of the convection instability. Prigogine tried to base his approach on thermodynamics, introducing concepts of entropy production and excess entropy production. As we now know, these concepts are, however, insufficient to deal with structure formation in such systems [87]. Based on a fundamental idea of A. Turing [139], Prigogine and Nicolis [114], see also [108], treated macroscopic pattern formation in a specific chemical reaction model. For more recent work see [107].

Because of the fundamental importance of thermodynamics, we elucidate its relationship to Synergetics more closely.

Thermodynamics (see for instance [19]) deals with systems in and out of thermal equilibrium. A central concept is entropy. In a closed system, it tends to its maximum value. Thermal equilibrium is characterized by the equipartition theorem: each degree of freedom has an average energy of  $1/2 kT$ ,  $k$  = Boltzmann constant,  $T$  absolute temperature. This may refer e. g. to gas atoms as well as to collective excitations in crystals. These systems are in thermal equilibrium with their surrounding (heatbaths, reservoirs).

Irreversible thermodynamics [51] treats systems which are not in thermal equilibrium but close to it. It mainly deals with transport and relaxation processes. A central concept is entropy production.

In the domains of physics, chemistry, biology, Synergetics deals with systems far from (thermal) equilibrium. This state is caused and maintained by an in- and outflow of matter, energy and / or information. This is achieved by a coupling of the “proper system” to heat baths (reservoirs) at different temperatures. The former concepts of thermodynamics, in particular the first and second law, are still valid for the total system (“proper” plus reservoirs), but no more sufficient to deal with the kinetics of the proper system. Now the central concept is growth and decay rates. In systems far from thermal equilibrium, collective modes are formed. One or several of them compete best for the external supply of matter, energy, information and grow at the expense of all other degrees of freedom (or modes). Thus the equipartition theorem is no more valid. In general, the behavior of the system is governed by few degrees of freedom (order parameters). Inci-



**Synergetics: Basic Concepts, Scheme 1**  
**Quantum optics, example laser (after [54])**

dentally, this “growth and competition” principle applies to a great variety of fields out of physics, chemistry and biology, where “modes” may not only be special physical structures, but may mean behavioral patterns, special functions etc. Quite often, a “mode” is initiated by a chance event (fluctuation). Clearly, a generalized Darwinian principle can be seen: The interplay between *mutations* (microscopic chance events) and *selection* (competition between macroscopic modes) leads to macroscopic patterns (structures) in the widest sense of the word.

In present days research, a new name is spreading, namely complexity or complexity theory. There seems to be no precise definition of this field available in the scientific community. Of what is known so far, we may conclude that this field has strong ties to the original field of

Synergetics in that it searches also for general principles but, in addition, it allows the collection or accumulation of knowledge on all kinds of complex systems, as is witnessed in the excellent Complexity Digest, weekly edited by Gottfried Mayer. What “complexity” eventually might mean is reflected by the present encyclopedia.

### Future Directions

Synergetics is surely not a closed scientific discipline but quite open to further research. On the one hand we may think of further applications of the principles of Synergetics that have been hitherto elaborated on, such as order parameters etc. Here a wide field of application is provided by robotics, construction of prostheses, automatic steering

of cars etc. On the other hand, new ideas to endow systems with self-organizing properties are needed, e.g. groups of mobile agents for the execution of specific tasks. First steps have been done for instance by Kornienko [81].

## Bibliography

Basically, the individual volumes of the Springer Series in Synergetics and selected volumes of the Springer Series on Complexity provide further information. Here, because of lack of space, only a few basic papers can be quoted.

The number of books, reviews and original papers on the topic of the present article is enormous. I have tried to achieve a fair balance between original contributions and reviews / books. Nevertheless, my selection must remain – to some extent – arbitrary and incomplete.

## Primary Literature

1. Abraham R, Marsden JE (1978) Foundations of mechanics. Benjamin / Cummings, Reading
2. Andronov A, Vitt A, Khaikin SE (1966) Theory of oscillators. Pergamon Press, London-Paris
3. Arnold VI (1963) Russ Math Surv 18:9
4. Arnold VI (1993) Dynamical systems VI singularity theory 1 v 6 (Encyclopaedia of Mathematical Sciences). Springer, Berlin
5. Arnold VI, Afrajmovich VS, Il'yashenko, Yu S, Shil'nikov Lf (1999) Bifurcation theory and catastrophe theory. Springer, Berlin
6. Aronson IS, Kramer L (2002) The world of the complex Ginzburg–Landau equation. Rev Mod Phys 74:99–143
7. Babloyantz A (1986) Molecules, dynamics and life: An introduction to Self-Organization of matter. Wiley, Indianapolis
8. Bartholomew DJ (1967) Stochastic models for social processes. Wiley, London
9. Belousov BP (1959) Sb Ref radats Med, Moscow
10. Bertalanffy L von (1950) Brit J Phil Sci 1:134
11. Bertalanffy von L (1953) Biophysik des Fließgleichgewichts. Vieweg, Braunschweig
12. Bodenschatz E, Pesch W, Ahlers G (2000) Annu Rev Fluid Mech 32:709
13. Bogoliubov NN, Mitropolsky YA (1961) Asymptotic methods in the theory of nonlinear oscillations. Hindustan Publ Corp, Delhi
14. Bonner JT, Barkley PS, Hall EM, Konjin TM, Mason JW, O'Keefe G, Wolfe PB (1972) Devel Biol 20:72
15. Bray CH (1921) J Am Chem Soc 43:1262
16. Busse FH (1972) J Fluid Mech 52:1–97
17. Bénard H (1900) Ann Chim Phys 7(23):62
18. Bénard H (1900) Rev Gén Sci Pures Appl 11:1261–1309
19. Callen HB (1960) Thermodynamics. Wiley, New York
20. Chandrasekhar S (1961) Hydrodynamic and hydromagnetic stability. Clarendon Press, Oxford
21. Chenciner A, Iooss G (1979) Arch Ration Mech Anal 69:109
22. Chow S-N, Hale JK (1982) Methods of bifurcation theory. Springer, Berlin
23. Collet P, Eckmann JP (1980) Iterated maps on the interval as dynamical system. Birkhäuser, Boston
24. Cross MC, Hohenberg P (1993) Rev Mod Phys 65:851
25. DeGiorgio V, Scully MO (1970) Phys Rev A 2:1170
26. Eigen M, Schuster P (1977) Naturwissenschaften 64:541, (1978) 65:7, 65:341
27. Elsasser W (1937) Phys Rev 52:987, (1968) Z Phys 171
28. Epstein IR, Pojman JA (1998) Introduction to nonlinear chemical dynamics. Oxford University Press, New York
29. Fantz M, Bestehorn M, Friedrich R, Haken H (1993) Phys Lett A 174:48–52
30. Feigenbaum MJ (1978) J Stat Phys 19:25
31. Fenstermacher RP, Swinney HL, Gollub JP (1979) J Fluid Mech 94:103
32. Field RJ, Korös E, Noyes RM (1972) Am Chem Soc 49:8649
33. Fife PC (1979) Mathematical aspects of reacting and diffusing systems. Springer, Berlin
34. Fisher G (1967) Measuring ambiguity. Am J Psychol 80:541–547
35. Floquet G (1883) Sur les équations différentielles linéaires à coefficients périodiques. Ann École Norm 2(12):47
36. Gardiner CW (1994) Handbook of stochastic methods. Springer Series in Synergetics, vol 13. Springer, Berlin
37. Gerisch G, Hess B (1974) Proc Nat Acad Sci (Wash) 71:2118
38. Gaiiotti DB, Steinacker R, Stel F (2007) Atmospheric convection. Research and Operational Forecasting Aspects, Cism International Centre for Mechanical Sciences Courses and Lectures. Springer, Wien
39. Gierer A, Meinhard H (1972) Kybernetik 12:30
40. Glansdorff P, Prigogine I (1971) Thermodynamic theory of structure, stability, and fluctuations. Wiley, New York
41. Gollup J, Benson SV (1979) In: Haken H (ed) Pattern Formation by Dynamic Systems and Pattern Recognition. Springer Series in Synergetics, vol 5. Springer, Berlin
42. Golubitsky M, Schaeffer D (1988) Singularities and groups in bifurcation theory I, vol 1. Springer, Berlin
43. Golubitsky M, Stewart I, Schaeffer D (1988) Singularities and groups in bifurcation theory, vol 2. Springer, Berlin
44. Graham R (1970) Quantum statistics of optical parametric oscillation. In: Kay SM, Maitland A (eds) Quantum Optics. Academic Press, New York
45. Graham R (1981) Z Phys B 40:149
46. Graham R, Haken H (1968) Z Phys 213:420 (1970) 235, 237:31,166
47. Graham R, Haken H (1971) Z Phys 248:289
48. Grossmann S, Thomae S (1977) Z Naturforsch 32A:1353
49. Guckenheimer J, Holmes P (1983) Nonlinear oscillations, dynamical systems, and bifurcations of vector fields. Springer, Berlin
50. Haake F (1973) In: Springer Tracts in Modern Physics, vol 66. Springer, Berlin, p 98
51. Haase R (1969) Thermodynamics of irreversible processes. Addison-Wesley, Reading
52. Hahn W (1967) Stability of motion. In: Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen, Bd 138. Springer, Berlin
53. Haken H (1964) Z Phys 181:96
54. Haken H (1970) In: Encyclopedia of Physics, vol XXV/2c: Laser Theory. Springer, Berlin
55. Haken H (1975) Phys Lett 53A:77
56. Haken H (1975) Z Phys B 21:105, B 22:69, B 23:388
57. Haken H (1979) Light, vol 1, Elements of Quantum Optics. North-Holland Physics Publishing, Amsterdam, New York



58. Haken H (1985) *Light*, vol 2, *Laser Light Dynamics*. North-Holland Physics Publishing, Amsterdam, New York
59. Haken H (1996) *Principles of brain functioning*. Springer, Berlin
60. Haken H (2000) *Information and Self-Organization*, 2nd edn. Springer, Berlin
61. Haken H (2004) *Synergetic computers and cognition*, 2nd edn. Springer, Berlin
62. Haken H (2004) *Synergetics: An introduction and advanced topics*. Springer, Berlin
63. Haken H, Graham R (1971) *Umschau* 6:191
64. Haken H, Kelso S, Bunz H (1985) *Biol Cybern* 51:347
65. Haken H, Wunderlin A (1982) *Z Phys B* 47:179
66. Hansch D (2002) *Evolution und Lebenskunst. Grundlagen der psychosynergetik. Ein Selbstmanagement-Lehrbuch*. Vandenhoeck und Ruprecht, Göttingen
67. Hopf E (1942) Abzweigung einer periodischen Lösung eines differentialsystems. *Berichte der Mathematisch-Physikalischen Klasse der Sächsischen Akademie der Wissenschaften zu Leipzig XCIV 1*, Leipzig
68. Hopf E (1948) *Commun Pure Appl Math* 1:303
69. Horsthemke W, Lefever R (1983) *Noise-Induced transitions*, Springer Series in Synergetics, vol 15. Springer, Berlin
70. Hoyle RG (2006) *Pattern formation*. Cambridge UP, Cambridge
71. looss G, Joseph DD (1980) *Elementary stability and bifurcation theory*. Springer, Berlin
72. Itô K (1969) *Stochastic processes*. Universitet Matematisk Institut, Aarhus
73. Jaynes ET (1957) *Phys Rev* 106:4, 620, *Phys Rev* 108:171
74. Jaynes ET (1967) In: *Delaware Seminar in the Foundations of Physics*. Springer, Berlin
75. Kadanoff LP, Götze W, Hamblen D, Hecht R, Lewis EAS, Palcan-skas VV, Rayl M, Swift J, Aspnes D, Kane J (1967) *Rev Mod Phys* 39:395
76. Kelley A (1967) In: Abraham R, Robbin J (eds) *Transversal Mappings and Flows*. Benjamin, New York
77. Kelso JAS (1981) *Bull Psychonomic Soc* 18:63
78. Kelso JAS (1995) *Dynamic patterns: The self-organization of brain and behavior*. MIT Press, Cambridge
79. Kernbach S (2008) *Structural self-organization in multi-agent and multi-robotic systems*, Thesis. Stuttgart University, Logos, Berlin
80. Kierhöfer, Hansjörg (2004) *Bifurcation theory, an introduction with applications to PDES*. Springer, Berlin
81. Kolmogorov AN (1954) *Dokl Akad Nauk USSR* 98:527
82. Kuhn TS (1996) *The structure of scientific revolutions*, 3rd edn. University of Chicago Press, Chicago
83. Kuramoto Y (1984) *Chemical oscillations, waves and turbulence*. Springer, Berlin
84. Kuznetsov, Yuri A (1995) *Elements of applied bifurcation theory*. Springer, Berlin
85. Köhler W (1920) *Die physischen gestalten in ruhe und im stationären zustand*. Vieweg, Braunschweig
86. Landau LD, Lifshitz IM (1959) In: *Course of Theoretical Physics*, vol 5, *Statistical Physics*. Pergamon Press, London-Paris
87. Landauer R (1975) *Phys Rev A* 12:636
88. Langevin P (1908) *Sur la théorie du mouvement brownien*. *CR Acad Sci Paris* 146:530
89. Lorenz EN (1963) *J Atmospheric Sci* 20:130, 20:448
90. Lyapunov AM (1906) *Sur la masse liquide homogène donnée d'un mouvement de rotation*. *Zap Acad Nauk St. Petersburg* 1:1
91. Ma, Tian, Wang, Shouhong (2005) *Bifurcation theory and applications*. World Scientific, Singapore
92. Manneville P (1990) *Dissipative structures and weak turbulence*. Academic Press, San Diego
93. Marx K (1987) *Analytische und numerische behandlung der zweiten instabilität beim Taylor-Problem der flüssigkeitsdynamik*, Thesis. Stuttgart University, Shaker, Aachen
94. May RM (1976) *Nature* 261:459
95. Meinhardt H (1982) *Models of biological pattern formation*. Academic, London
96. Meinhardt H (1990) *The beauty of sea shells*. Springer, Berlin
97. Meystre P, Sargent M (1990) *Elements of quantum optics*. Springer, Berlin
98. Mikhailov AS (1993) *Foundations of synergetics*. In: *Distributed Active Systems, II* (with AY Loskutov): *Complex Patterns*. Springer, Berlin
99. Mills DL (1991) *Nonlinear optics basic concepts*. Springer, Berlin
100. Moser J (1967) *Math Ann* 169:136
101. Murdock J (2002) *Normal forms and unfoldings for local dynamical systems*. Springer, New York
102. Murray JD (1989) *Mathematical biology*, 2nd edn 1993, 3rd edn 2002/2003. Springer, Berlin
103. Nayfeh, Ali H (1993) *Method of normal forms*. Wiley, New York
104. Nekorkin VI, Velarde MG (2002) *Synergetic phenomena in active lattices. Patterns, waves, solitons, chaos*. Springer, Berlin
105. Newell AC, Whitehead JA (1969) *J Fluid Mech* 38:279
106. Newhouse S, Ruelle D, Takens F (1978) *Commun Math Phys* 64:35
107. Nicolis G (1995) *Introduction to nonlinear science*. Cambridge University Press, Cambridge
108. Nicolis G, Prigogine I (1977) *Self-organization in non-equilibrium systems*. Wiley, New York
109. Pauli H (1928) *Probleme der modernen physik*. In: P Debye (ed) *Festschrift zum 60. Geburtstag A Sommerfelds*. Hirzel, Leipzig
110. Pismen LM (1999) *Vortices in nonlinear fields*. Clarendon Press, Oxford
111. Pismen LM (2006) *Patterns and interfaces in dissipative dynamics*. Springer, Berlin
112. Pliss VA (1964) *Izv Akad Nauk SSSR, Mat Ser* 28:1297
113. Poincaré H (1960) *Les methods nouvelles de la mécanique céleste*. Gauthier-Villars, Paris. Reprint 1892/99 Dover Publ, New York
114. Prigogine I, Nicolis G (1967) *J Chem Phys* 46:3542
115. Rabinovich MI, Eversky AB, Weidmann PD (2000) *The dynamics of patterns*. World Scientific, Singapore
116. Risken H (1965) *Z Phys* 186:85
117. Risken H (1989) *The Fokker Planck eq*. Springer, Berlin
118. Ruelle D, Takens F (1971) *Commun Math Phys* 20:167
119. Sargent M, Scully MO, Lamb WE (1974) *Laser physics*. Addison-Wesley, Reading
120. Sattinger DH (1980) *Group theoretic methods in bifurcation theory*. *Lecture Notes Math*, vol 762. Springer, Berlin
121. Schawlow AL, Townes CH (1958) *Phys Rev* 112:1940
122. Schiepek G (1999) *Die grundlagen der systemischen therapie. Theorie-Praxis-Forschung*. Vandenhoeck und Ruprecht, Göttingen



123. Schleich WP (2001) Quantum optics in phase space. Wiley-VCH, Weinheim
124. Schmidt E (1908) Zur theorie der linearen und nichtlinearen integralgleichungen, Teil 3. Math Annalen 65:370
125. Schöll E (2001) Nonlinear spatio-temporal dynamics and chaos in semiconductors. Cambridge University Press, Cambridge
126. Scully M, Lamb WE (1967) Phys Rev 159:208 (1968) 166:246
127. Segel LA (1969) J Fluid Mech 38:203
128. Sell GR (1979) Arch Ration Mech Anal 69:199
129. Shannon CE, Weaver W (1949) The mathematical theory of communication. Univ of Illin Press, Urbana
130. Smale S (1967) Bull AMS 73:747
131. Sounders PT (1980) An introduction to catastrophe theory. Cambridge University Press, Cambridge
132. Sparrow CT (1982) The Lorenz equations: bifurcations, chaos and strange attractors. Springer, Berlin
133. Staliunas K, Sanchez-Morcillo V, Gaul LJ (2003) Transverse patterns in nonlinear optical resonators. Springer, Berlin
134. Stratonovich RL (1963) Topics in the theory of random noise, vol 1, 1967 vol II. Gordon Breach, New York-London
135. Swift J, Hohenberg PC (1977) Phys Rev A 15:319
136. Swinney HL, Gollub JP (eds) (1981) Hydrodynamic instabilities and the transition to turbulence, Topics Appl Phys, vol 45. Springer, Berlin
137. Taylor GI (1923) Phil Trans R Soc Lond A 223:289
138. Thom R (1975) Structural stability and morphogenesis. Benjamin, Reading
139. Turing AM (1952) Phil Trans R Soc Lond B 237:37
140. Vavilin VA, Zhabotinsky AM, Yaguzhinsky LS (1967) Oscillatory processes in biological and chemical systems. Moscow Science Publ, Moscow, p 181
141. Walgraef D (1997) Spatio-temporal pattern formation. Springer, New York
142. Walls DF, Milburn GJ (1994) Quantum optics. Springer, Berlin
143. Weidlich W (2000) Sociodynamics. A systematic approach to mathematical modelling in the social sciences. Harwood Academic Publishers, Amsterdam
144. Weidlich W, Haake F (1965) Z Physik 186:203
145. Wigner EP (1932) Phys Rev 40:749
146. Wilson KG, Kogut J (1974) Phys Rep 12 C:75
147. Wolpert L (1969) J Theor Biol 25:1
148. Wunderlin A, Haken (1981) Z Phys B 44:135
149. Zaikin AN, Zhabotinsky AM (1970) Nature 225:535

## Books and Reviews

Springer Series in Synergetics. Founded by Haken H (1977) vols 1–ca 100. Springer, Berlin

## Synergetics, Introduction to

HERMANN HAKEN

Institut für Theoretische Physik, Universität Stuttgart,  
Stuttgart, Germany

Synergetics (Greek **synergeon**: science of cooperation) is an interdisciplinary field of research that deals with

the behavior of complex systems, i. e. systems composed of many individual parts that in general may produce complex behavior. The systems considered may belong to a great variety of fields, ranging from natural sciences with physics, chemistry, biology, through medicine and psychology to economy, ecology, management theory etc. The central topic of the synergetics enterprise is this: By means of their cooperation, the individual parts of a system can spontaneously produce structures, i. e. special arrangements between the elements or specially coordinated behavior without specific steering from the outside, i. e. without the interference of an external agent. The physical systems, studied by synergetics, are away from thermal equilibrium. The structures may be spatial patterns, be it regular or irregular, temporal patterns, i. e. all kinds of self-sustained oscillations, ranging from harmonic oscillations to chaotic behavior, spatio-temporal patterns, or functional patterns such as produced, for instance, by the brain of humans or animals, giving rise to specific behavior. The basic question synergetics asks is this: Are there general principles that govern the self-organization of systems, irrespective of the nature of the individual parts? When this question was posed some 40 years ago by H. Haken, the problem put forward seemed to be absurd in view of the great variety of possible elements. However, this research program has turned out to be quite successful as witnessed, for instance by the Springer Series in Synergetics, as well as by an independent book series in Russian. The price to be paid for approaching this goal is as follows:

1. Focus the study on those situations where the **macroscopic state** of a system changes **qualitatively**.
2. Start from comparatively simple systems that are provided either by systems in physics (e. g. lasers or fluids) or model systems dealing with self-organization processes in a variety of fields, such as biology.

As has become evident over the past few decades, self-organization phenomena occur in a vast variety of fields whose presentation actually would fill a whole encyclopedia, which would have led to an encyclopedia within an encyclopedia. Indeed the reader him- or herself may find numerous examples in his/her own field of research. In view of this situation, I felt it would be wise to present to the reader a few basic concepts and only a few prototypical examples. It is, indeed, an outstanding fact that few concepts allow us to cover a great variety of self-organization phenomena from a unifying point of view. Among these concepts are stability, instability, control parameters, order parameters and the slaving principle. Future research

will have to add more concepts as has already happened, for instance, in chaos theory.

As the reader will note, from the theoretical point of view, synergetics has become a meeting place between bifurcation theory and mathematical theories related to it, the theory of stochastic processes, and phase transition theory, actually in the sense of the Landau theory. The mathematical tools employed are generalized Langevin equations and the Fokker–Planck equation, and to some extent also the density matrix equation. In physics, it has become possible to start from first principles. For instance in quantum optics, based on Heisenberg equations of motion for operators, the coherence properties of laser light (and other non-linear optical devices) were derived in every detail (its explicit presentation would require, however, a whole handbook article, so that the reader is referred to the original literature). Fluid dynamics was chosen as a highly illustrative and prototypical example of pattern formation in systems away from thermal equilibrium. When a fluid is energetically excited, with an increasing degree of excitation it may run through a hierarchy of spatio-temporal patterns. At comparatively low excitation levels, well defined patterns evolve. While regular patterns stand in the foreground of the article ► [Fluid Dynamics, Pattern Formation](#) by Bestehorn, the article by Pismen ► [Patterns and Interfaces in Dissipative Dynamics](#) emphasizes defects and interfaces. At higher excitation levels, we reach the field of turbulence, which even after the ground-breaking work by Kolmogorov is still a hot subject of research. Important steps are done here, e. g. by the study of the dynamics of vortices. The work by Friedrich and Peinke, authors of the article ► [Fluid Dynamics, Turbulence](#) comprises also the Fokker–Planck equation approach. Actually, in the introductory article ► [Synergetics: Basic Concepts](#) by Haken, a method using the concept of information (entropy) is outlined how to derive a Fokker–Planck equation from measured data. Friedrich (private communication) and Peinke found a more direct access, the results of which are briefly outlined in the article by Friedrich and Peinke.

A quite modern line of research on complex systems follows up the method of the Fokker–Planck equation and especially of **non-linear** Fokker–Planck equations. This will be outlined in the article ► [Linear and Non-linear Fokker–Planck Equations](#) by Frank.

So far, all these articles deal with “hard” science, especially physics. As it has turned out over past decades, the general concepts of synergetics have great potentialities in fields often called “soft” science such as movement science, medicine, and even psychology and psychiatry. Actually, as the following articles will show, basic concepts of

synergetics help to convert soft science into hard science. The article by Fuchs and Kelso ► [Movement Coordination](#) deals with movements of humans (and animals). But though these “systems” are highly complex, large classes of transitions between movement coordination can be theoretically and experimentally treated in great detail, even of a prototypical character.

While these approaches might be called “macroscopic”, the article by Tass, Popovych and Hauptmann ► [Brain Pacemaker](#) penetrates into the microscopic level of brain functions, namely by modeling the collective behavior of neurons and their reactions to external interventions. This paper is remarkable, because it shows a new aspect of the research on coupled nonlinear oscillators. Whereas so far the problem of how such systems synchronize, was in the foreground of research, now for medical reasons, namely to fight Parkinsons’s disease, the problem arises how to desynchronize such a system by appropriate interventions.

The mathematical theory of synergetics or perhaps in a wider sense dynamical systems theory with concepts such as attractors, allows one to cast earlier concepts developed by Gestalt theory into clear-cut mathematical models. The order parameters just play the role of Gestalt. Synergetics may be considered as a theory of indirect control of systems, in that it provides psychiatrists with insights for how to intervene with their clients. A comprehensive article, dealing with these aspects, is given in ► [Self-Organization in Clinical Psychology](#) by Schiepek and Perltz. This article shows how methods of mathematical analysis are penetrating more and more into the field of psychiatry, and allowing doctors to monitor the mental and behavioral state of a patient.

In conclusion, in particular with respect to biological systems (but not exclusively), the following remarks are in order:

1. While in physical systems, at least those considered in the article on synergetics, the control parameters are fixed from the outside, in biological systems the control parameters are in general produced by the system itself. They are the slowly varying variables in this case. When their dynamics is admitted, specific phenomena may occur, such as the recognition of ambivalent figures (switching between two or several percepts).
2. In biological and some physical processes, pattern formation (morphogenesis) is a two- or multi-step process. In the first step **dynamic structures** are formed. In a second step these dynamic patterns are transformed into solid patterns. A nice example is provided by radiolarians in which eventually dynamic concentration

patterns of chemicals are transferred into skeletons by means of calcification. The exploration of such two- or multi-step processes is, as it seems, just at its beginning.

In the article ► [Intentionality: A Naturalization Proposal on the Basis of Complex Dynamical Systems](#), W. Tschacher shows how mental processes can be linked to material processes where he tackles the longstanding problem of intentionality and relates it to a gradient dynamics which can be formalized by concepts of synergetics.

Last but not least, Synergetics as a theory dealing with systems composed of many individual parts deals with the interrelation between human individuals and their ability to form specific social groups, viewed as a process of self-organization. An outstanding example is provided by Juval Portugali's article ► [Self-Organization and the City](#).

## System Dynamics, Analytical Methods for Structural Dominance Analysis in

CHRISTIAN ERIK KAMPMANN<sup>1</sup>, ROGELIO OLIVA<sup>2</sup>

<sup>1</sup> Department of Innovation and Organizational Economics, Copenhagen Business School, Copenhagen, Denmark

<sup>2</sup> Mays Business School, Texas A&M University, College Station, USA

### Article Outline

Glossary

Definition of the Subject

Introduction

Characterizing Linear and Nonlinear Systems

Traditional Control Theory Approaches

Pathway Participation Metrics

Eigenvalue Elasticity Analysis

Eigenvectors and Dynamic Decomposition Weights (DDW)

Future Directions

Bibliography

### Glossary

**Behavior mode** The traditional meaning of the term is the qualitative nature of the observed system behavior, such as damped or expanding oscillations, overshoot and collapse, exponential growth or adjustment to equilibrium, or limit cycles. In linear systems theory, the term has a more specific meaning, cf. the explanation for eigenvalues.

**Boode plot (phase and gain plot)** A tool used in classical control theory to characterize the frequency response, i. e., the amplification  $A$  and phase shift  $\phi$  in the system output variable of interest  $x(t) = A \sin(\omega t + \phi)$  compared to the input variable  $u(t) = \sin(\omega t)$ , as a function of the frequency  $\omega$  of the input.

**Chaos** A type of behavior exhibited by nonlinear systems that appears to be approximately periodic but with a seemingly random element. A hallmark of chaotic behavior is that it is sensitive to initial conditions.

**Dominant structure** A general term for the feedback loops (or possibly external driving forces) that are "most important" in generating a behavior pattern of interest. In nonlinear models, particularly single-transient models, there is frequently a shift in structural dominance, i. e. in the strength and significance of certain feedback loops.

**Dynamic decomposition weights (DDW)** An application of Eigenvector Elasticity Analysis (EVA) that focuses on how parameter changes influence the relative weights (DDW's) of the system behavior modes in a particular variable.

**Eigenvalue** An eigenvalue for a square matrix  $A$  is a value  $\lambda$  for which the equation  $A\mathbf{r} = \lambda\mathbf{r}$  has a non-zero solution  $\mathbf{r} \neq \mathbf{0}$ . The column vector  $\mathbf{r}$  is called the (right) eigenvector corresponding to the eigenvalue  $\lambda$ . The eigenvalues and eigenvectors determine the behavior modes (components) in the solution to the linear dynamical system  $\dot{\mathbf{x}} = A\mathbf{x}$ . A real eigenvalue  $\lambda$  leads to an exponential behavior mode  $\exp(\lambda t)$  while a complex eigenvalue  $\lambda = \tau \pm i\omega$  leads to oscillatory behavior modes  $\exp(\tau t) \sin(\omega t + \phi)$ . The eigenvectors determine the weight, or the degree to which a particular behavior mode is expressed in a particular system variable.

**Eigenvalue elasticity analysis (EEA)** A method of analyzing the significance of a structural element, say a loop or a link in the model with a gain  $g$ , in terms of its marginal effect upon the eigenvalues  $\lambda$  of the system. There are several such measures, such as the *influence measure*  $\partial\lambda/\partial g \cdot g$ , the *elasticity*  $\partial\lambda/\partial g \cdot (g/\lambda)$ , or, in the case of complex-valued eigenvalues, the effect upon the damping ratio, natural frequency, damping time, etc., as illustrated in Fig. 7. See also *Loop Eigenvalue Elasticity Analysis (LEEA)*.

**Eigenvector** See explanation for *Eigenvalue*.

**Eigenvector elasticity analysis (EVA)** A complement to Eigenvalue Elasticity Analysis (EEA) that looks explicitly at the expression or relative weight of each behavior mode in each system variable. These weights are related to the eigenvectors of the system matrix.

**Frequency domain** A term used to describe the analysis of signals with respect to frequency. While a time domain graph shows the behavior of the signal over time, the frequency domain graphs shows how much of the signal variance lies within each given frequency band.

**Independent loop set (ILS)** Although the number of feedback loops in a model can be very large (theoretically astronomically large), there is a much smaller *independent loop set* that can be considered independent structural elements. For a strongly connected system (where any pair of variables are connected via causal chain in both directions) with  $N$  links and  $n$  variables, there are exactly  $N - n + 1$  independent loops. Simple algorithms exist for constructing independent loop sets, in particular *Shortest Independent Loop Sets (SILS)*. See also explanation for *Loop Eigenvalue Elasticity Analysis (LEEA)*.

**Linear dynamical system** A system where the rates  $\dot{\mathbf{x}} = (dx_1/dt, \dots, dx_n/dt)$  are a linear function of the state variables  $\mathbf{x} = (x_1, \dots, x_n)$  and exogenous or control variables  $\mathbf{u} = (u_1, \dots, u_p)$ , expressed by the equation  $\dot{\mathbf{x}} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{u}$  where  $\mathbf{A}$  is an  $n \times n$  matrix and  $\mathbf{B}$  is an  $n \times p$  matrix. Unlike nonlinear systems of the general form  $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u})$ , linear systems have analytical solutions based on the eigenvalues and eigenvectors of the matrix  $\mathbf{A}$  (cf. explanation for *Eigenvalues*).

**Linear systems theory** The mathematical theory of *linear dynamical systems*.

**Loop eigenvalue elasticity analysis (LEEA)** A form of eigenvalue elasticity analysis (EEA) that uses graph theory to express structural changes in terms of change in the strength of individual feedback loops. *Independent* loops can be assigned individual (loop) eigenvalue elasticities or influence measures just like other structural elements (see explanation for *Eigenvalue Elasticity Analysis (EEA)* and *Independent Loop Set (ILS)*).

**Model simplification approach** A way of attributing dynamic behavior to particular pieces of structure by replacing the full model with a simplified structure. See also *Structure contribution approach*.

**Nonlinear systems** Systems of the form  $\dot{\mathbf{x}} = \mathbf{f}(\mathbf{x}, \mathbf{u})$  where  $\mathbf{f}$  is a nonlinear function. See explanation for *Linear dynamical systems*.

**Pathway participation metric** A measure that decomposes the curvature ( $\ddot{x} = d^2x/dt^2$ ) of a variable  $x_i$  into the individual driving components,  $\ddot{x}_i = \sum_j \partial \dot{x}_i / \partial x_j \cdot \dot{x}_j$ . By considering the sign of the curvature relative to the slope, i. e.,  $\ddot{x}/\dot{x}$ , one may define behavior as (apparently) dominated by positive  $\ddot{x}/\dot{x} > 0$  or negative  $\ddot{x}/\dot{x} < 0$  feedback loops. The component

(pathway) with the largest absolute value and the same sign as  $\ddot{x}/\dot{x}$  is then defined as the dominant structure.

**Quasilinear models** Models that are almost linear in structure around the operating point of interest so that they may be well approximated by a linear model.

**Quasiperiodic behavior** A behavior that is a sum of oscillations of incommensurate frequencies so that the system never returns to exactly the same point (which would be the case for periodic behavior).

**Shortest independent loop set (SILS)** An *Independent Loop Set (ILS)* that consists of the shortest possible loops (in terms of the number of nodes and links in each loop). Since the choice of ILS is far from unique, an SILS provides a more focused choice of loops, which are typically also easier to interpret due to their short length.

**Single-transient models** Models where the behavior of interest is the transition toward an equilibrium or constant growth rate. Models are typically nonlinear, exhibit patterns such as smooth transition, overshoot and collapse, growth, or stagnation.

**Structure contribution approach** A way of linking model structure to dynamic behavior by considering how individual pieces of structure (feedback loops or subsystems) contribute to the behavior pattern of interest by turning the structure on or off (in traditional simulation experiments) or by considering the marginal effect of small changes in structure (the eigenvalue approach). See also *Model simplification approach*.

## Definition of the Subject

The link between system structure and dynamic behavior is one of the defining elements in the system dynamics paradigm, yet it is only recently that systematic, mathematically rigorous methods for exploring this link have started to become available. In a sense, a simulation model can be viewed as an explicit and consistent theory of the behavior it exhibits. Although this point of view has certain merits, not least the fact that it lifts the discussion from outcomes to causes of these outcomes and from events to underlying structure [11,59], we are concerned here with a more compact explanation of the system's behavior. In fact, most system dynamics modeling projects report their results in terms of simpler explanations of the observed results, typically in terms of dominant feedback loops that produce the salient features of the behavior.

Most often, dominant structure is thought of in terms of feedback loops and, occasionally, external driving forces to the system. For simple systems with relatively few vari-

ables it is usually easy to use intuition and trial and error simulation experiments to explain the dynamic behavior as resulting from particular feedback loops. In larger systems, this method becomes increasingly difficult and the risk of incorrect explanations rises accordingly. There is a need, therefore, for analytical methods that provide some consistency and rigor to this process.

These analytical tools are important to the practitioner because the structure-behavior link is the key to finding leverage points for policy initiatives. And they are important to the theorist because a system dynamics theory of a particular phenomenon is an account of how certain feedback loops cause certain dynamic patterns of behavior to appear. The qualitative understanding of the model behavior is often at least as important as the particular numerical predictions obtained, even in applied studies. Yet the rigor of such an account depends directly on the rigor with which structure-behavior link can be made in a given model.

The classical disciplines of linear systems theory and control engineering have provided a set of concepts and tools, particularly system eigenvalues and eigenvectors, that can also be applied under many circumstances to the nonlinear models found in system dynamics, not as a complete theory but as a pragmatic aid. This article reviews the recent advances in analytical tools based on linear systems theory and discusses its future potential for the both the system dynamics practitioner and the theorist.

Though we strongly believe in the utility of these methods, it is important to realize that advances in nonlinear dynamics and complexity theory in recent decades have shown that it is not possible to construct a complete theory of dominant structure because nonlinear systems are capable of exceedingly complex and intricate behavior that is impossible to predict without actually simulating the system. Furthermore, applications of graph theory to system dynamics models have revealed that the concept of feedback loops has some inherent problems and limitations because there are potentially many different loop descriptions of the same system (see [28,40]). Thus, the analytical tools should be viewed as pragmatic aids to model analysis that can guide the modeler's intuition, rather than universal methods that provide automatic answers.

We first provide a brief historical introduction to the different ways scholars have thought about the notion of dominant structure, including an example of the traditional approach to structural analysis. In the next section we present the formal mathematical representation of linear and nonlinear systems and how one may describe the dynamic behavior in terms of behavior modes and system eigenvalues. In the four following sections we present al-

ternative approaches to performing this analysis. We conclude with a summary of the current state of research and a discussion of future directions.

## Introduction

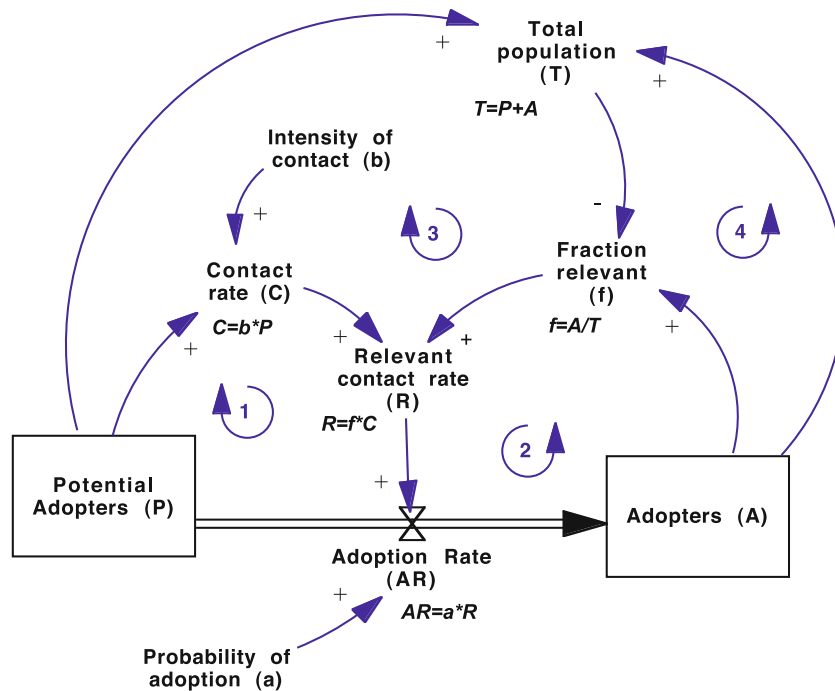
Understanding model behavior is closely related to the process of model testing and validation, for which there is a well-established tradition and an extensive literature in the field (e. g., [2,10,17,36,46,47]). Indeed there is no sharp line between model building, testing, validation, and analysis – in practice, the analyst undertakes all these processes simultaneously [17].

Of particular concern is whether one can identify pieces of structure that are in some sense “important” in generating the observed behavior of interest. Traditionally, system dynamics analysts have relied on trial-and-error simulation to discover these structures, by changing parameter values or switching individual links and feedback loops on and off. The tradition is well developed and includes a set of principles for partial model formulation and testing based the organizational theory of bounded rationality [27,36].

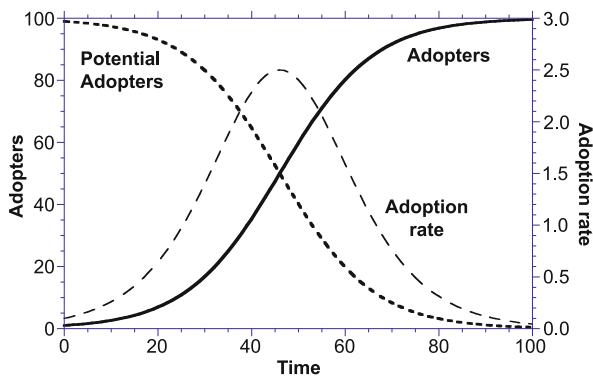
The intuition guiding this effort often relies on simple feedback systems with one or a few state variables, where the behavior is fully documented and understood. In particular, the modeler uses well-understood “generic structures” that seem to appear again and again in system dynamics models, such as “overshoot and carrying capacity collapse”, “drifting goal structure”, etc. (see [30,56,60] for an account of these structures). Clearly such structures can be a useful aid to understanding if the model is sufficiently simple to allow such simple structures to be identified.

A simple example of a generic structure is the classical model of diffusion, sometimes known as the Bass model ([3], see also Chapter 9 in [59]). The model structure is illustrated in Fig. 1, and the resulting behavior, an s-shaped growth curve, is illustrated in Fig. 2. The idea behind the model is that the adoption of a new technology is driven by the number of users that have already adopted it, through a word-of-mouth effect. One may interpret the s-shaped behavior as the interaction of two feedback loops, namely loop no. 2, the positive “word-of-mouth”, and loop no. 1, the negative “exhaustion” loop (see Fig. 1). In the beginning, the positive loop dominates, leading to exponential growth in the number of adopters. Later, however, the negative loop gains strength, and the behavior shifts to an exponential adjustment toward the eventual market saturation. Thus, the traditional feedback loop analysis helps give an intuitive understanding of the dynamics of the model.





System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 1  
The Bass model of diffusion



System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 2  
Behavior of the Bass model

In large-scale models with perhaps hundreds of state variables, however, the traditional approach shows significant limitations. In practice, model building and analysis is often done using a “nested” partial model testing approach where one goes from the level of small pieces of structure to entire subsystems of the model, with frequent re-use of known formulations and partial models. Although this approach does carry a long way, it can be very difficult to discover feedback mechanisms that transcend model substructures in ways not anticipated by the modeler in the

original dynamic hypothesis. Thus, there is a danger that observed behavior is falsely attributed to certain feedback mechanisms when in fact another set of feedbacks is driving the outcome. Likewise, one may make false inferences about how a particular feedback mechanism modifies the behavior, e. g., whether it attenuates or amplifies a particular oscillation.

Modern software packages can run extensive tests for sensitivity and “reality checks” where a large number of parameters are varied simultaneously [44]. This is clearly a significant improvement over “manual” trial and error methods, particularly when these methods are combined with statistical inference methods such as Kalman Filtering or Monte Carlo maximum likelihood estimation [6,8,39,43,45,55]. A variant of this approach involves using statistical experimental design and correlation methods to screen for significant model structure (parameters), as suggested by Ford and Flynn [9]. Indeed, the prospects of marrying such methods with modern search and optimization methods like classifier systems [26] or genetic algorithms [19] seem very promising. However, these methods are more addressing issues in estimation, validation and testing than inferences about or understanding how (dominant) structure is causing behavior.

Richardson [47] suggested a taxonomy of approaches to the notion of dominant structure, where he distin-

guishes along three dimensions, namely linear vs. nonlinear systems, model reduction vs. loop contribution, and the characterization of behavior in terms of time graphs vs. eigenvalues or frequency response. Of these, the distinction between model reduction and loop contribution is the most important.

In the model reduction approach, the idea is to replace a large complicated model with a simplified smaller model that captures the “essence” of the dynamics. A good example of this is Sterman’s simple model of the economic long wave [58], which was distilled from the much larger System Dynamics National Model [18]. Eberlein [5,7] attempted to tackle model simplification in a systematic way in linear systems by focusing on retaining specific behavior modes. In large part his results were negative: it is generally not possible to build simpler models that reproduce the salient behavior without sacrificing either the accuracy of the behavior or the ability to relate the simplified model variables to those in the full model. It is fair to say that this line of inquiry has largely been abandoned as a result. Extracting the “essence” of a model remains an art more than a science.

The focus here will be on Richardson’s second category, the loop contribution or, more generally, the *structure contribution* approach. It reflects the intuitive idea that if one removes the element under consideration, e.g. by weakening a link or switching off a feedback loop, and the behavior then “disappears”, one would say that the element in some sense “causes” the observed behavior.

This notion underlies the traditional trial-and-error simulation approach, sometimes supplemented with methods from the classical control engineering, which focuses on how structural elements modify the behavior of the system, viewed in terms of the frequency response. Typically, the method works “backwards” by starting with simple feedback systems of single loops and then considering the marginal effect of adding links and loops. We discuss this approach in Sect. “[Traditional Control Theory Approaches](#)” below.

If, instead, one considers marginal (infinitesimal) changes in structure, e.g. in the strength of a particular link, it is possible to derive rigorous analytical results for the resulting change in behavior expressed as the eigenvalues of the linearized model. One would then say that if a change in a system element has a relatively large effect upon the behavior pattern of interest, this element is “significant” in “causing” the behavior. This is what underlies the *eigenvalue elasticity* and *eigenvector* approaches discussed in Sects. “[Eigenvalue Elasticity Analysis](#)”, “[Eigenvectors and Dynamic Decomposition Weights \(DDW\)](#)”. The marginal and experimental approaches may supple-

ment each other well, where a marginal analysis may identify elements that can then be tested experimentally for their significance.

Unlike the traditional control method and the eigenvalue method that work in the structural and *frequency domain*, the *pathway participation* method (PPM) relates directly to the time path of particular system variables and is more concerned with the qualitative nature of the time path, expressed in terms of signs of the slope (whether growing or declining) and curvature (whether convex or concave) than with numerical measures of degree of influence. Briefly stated, the PPM traces the causal links in the variables influencing the system variable in question and then identifies the most important chain of links. We discuss this method in Sect. “[Pathway Participation Metrics](#)”.

Common to the approaches discussed here is that they all build upon a precise mathematical characterization of the system behavior. In the next section, we demonstrate how the concepts from linear systems theory may be used to give a precise characterization of behavior in terms of component *behavior modes*.

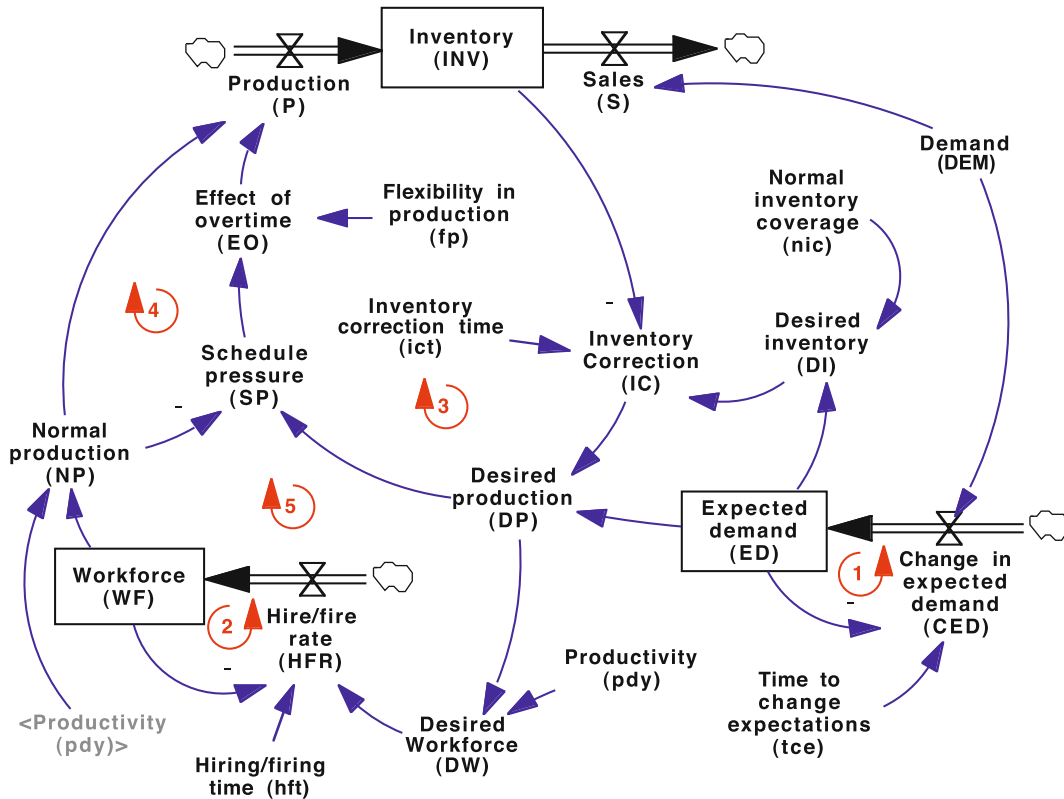
## Characterizing Linear and Nonlinear Systems

A system dynamics model can be represented mathematically as a set of ordinary differential equations

$$\frac{dx(t)}{dt} \equiv \dot{x}(t) = f(x(t), u(t)) , \quad (1)$$

where  $x(t)$  is a (column) vector of  $n$  state variables (levels)  $(x_1(t), \dots, x_n(t))$ ,  $u(t)$  is a column vector of  $p$  exogenous variables or control variables  $(u_1(t), \dots, u_p(t))$ ,  $f(\cdot)$  is a corresponding vector function, and  $t$  is simulated time. In this paper, we restrict our attention to the state variables (levels) of the model for notational convenience, ignoring the auxiliary variables. Mathematically, a model can always be brought to the *reduced* form (1), but in practice, the auxiliary variables give a more intuitive account of the analysis. Likewise, we do not consider time-varying systems (where time  $t$  enters as an explicit argument in the function  $f$ ), since these can usually be accommodated by an appropriate definition of the exogenous variables  $u$ . In general,  $f$  is a nonlinear function of its arguments, and we speak of a *nonlinear* system. Conversely, if  $f$  is a linear function, we speak of a *linear* system.

Figure 3 and Table 1 show a well-known example, the inventory–workforce model. It has three state variables, Inventory (INV), Workforce (WF), and Expected Demand



System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 3  
Flow diagram of the inventory workforce model

(ED), and one exogenous variable, Demand (DEM), i. e.,

$$\mathbf{x}(t) = \begin{pmatrix} \text{INV} \\ \text{WF} \\ \text{ED} \end{pmatrix}; \quad \mathbf{u}(t) = (\text{DEM}), \quad (2)$$

and the function  $\mathbf{f}$  is determined by the equations in Table 1.

Given the model structure (1), knowledge of the initial conditions  $\mathbf{x}(0)$ , and the path of the input variables  $\mathbf{u}(t)$ , the behavior of the model is completely determined. It is in this sense that the model structure (1) constitutes a “theory” of the time behavior  $\mathbf{x}(t)$ , as mentioned in the introduction. Yet, we are interested in methods that yield a more compact explanation, short of having to simulate the entire model structure.

It turns out that in its ultimate form, this dream is beyond reach: Since the days of Henri Poincaré, mathematicians have known that it is impossible to find general analytical solutions to nonlinear systems. Furthermore, the development of nonlinear dynamics and chaos theory has proven that such systems, even when they have very few state variables, can produce highly complex and intricate behavior that goes beyond general analytic methods

(e. g., [42,48]). Thus, we will never find a final general theory where we can infer the behavior of the system directly from its structure; instead, we will always have to rely on simulation to discover the dynamics implied by the structure. (This is not to say that no general analytical results exist in nonlinear systems. The field of chaos theory has uncovered a number of universal features, e. g., relating to the transition from periodic or quasi-periodic behavior to chaos, where the transitions show both qualitative and quantitative similarities that are independent of the specific forms of the model equations (see, e. g., [42]). However, these universal features relate to specific situations such as period-doubling or intermittency routes to chaos).

The best we can hope for, therefore, is a set of tools that will guide intuition and help identify *dominant structure* in the model. By dominant structure we mean particular feedback loops that are in some sense “important” in shaping the behavior of interest. To the extent that we can identify such dominant structures, we may say that we have found a “theory” of the observed behavior.

Although the term “behavior” may appear rather loose, experience and reflection tells us that there is a limited number, perhaps a dozen or so, of relevant behav-

**System Dynamics, Analytical Methods for Structural Dominance Analysis in, Table 1**  
**Equations of the inventory workforce model**

| Equation                                     | Name                        | Units                |
|----------------------------------------------|-----------------------------|----------------------|
| $d/dt(ED) = (DEM - ED)/tce, ED_0 = DEM * df$ | Expected demand             | [Units/Month]        |
| $d/dt(INV) = P - S, INV_0 = DI$              | Inventory                   | [Units]              |
| $d/dt(WF) = HFR, WF_0 = DWF$                 | Workforce                   | [Workers]            |
| $S = DEM$                                    | Shipments                   | [Units/Month]        |
| $P = NP * EO$                                | Production                  | [Units/Month]        |
| $EO = fp * (1 - (1 - 1/fp)SP)$               | Effect of overtime          | [Dimensionless]      |
| $NP = WF * pdy$                              | Normal production           | [Units/Month]        |
| $DI = ED * nic$                              | Desired inventory           | [Units]              |
| $SP = DP/NP$                                 | Schedule pressure           | [Dimensionless]      |
| $DEM = 1(\text{Exogenous})$                  | Demand                      | [Units/Month]        |
| $DP = ED + IC$                               | Desired production          | [Units/Month]        |
| $IC = (DI - INV)/tci$                        | Inventory correction        | [Units/Month]        |
| $HFR = (DWF - WF)/hft$                       | Hire/fire rate              | [Workers/Month]      |
| $DWF = ED/pdy$                               | Desired workforce           | [Workers]            |
| $hft = 5$                                    | Hire/fire time              | [Month]              |
| $pdy = 1$                                    | Productivity                | [Units/Month/Worker] |
| $tce = 4$                                    | Time to change expectations | [Month]              |
| $fp = 1.05$                                  | Flexibility in production   | [Dimensionless]      |
| $nic = 3$                                    | Normal inventory coverage   | [Months]             |
| $ict = 2$                                    | Inventory correction time   | [Months]             |
| $df = 0.5$                                   | Disequilibrium fraction     | [Dimensionless]      |

ior patterns that dynamical systems can exhibit. Some of these behaviors, like exponential growth, exponential adjustment, and damped or expanding oscillations, are typical of linear systems. Others, like limit cycles, quasiperiodic motion, mode-locking, and chaos, can only be exhibited by nonlinear systems.

Common to the approaches considered in this paper is that they are based on tools from linear systems theory, i. e., they approximate the nonlinear model (1) with a linearized version, using first-order Taylor expansion around some operating point  $\mathbf{x}_0, \mathbf{u}_0$ , i. e.,

$$\dot{\mathbf{x}}(t) \approx \mathbf{f}(\mathbf{x}_0, \mathbf{u}_0) + \frac{\partial \mathbf{f}}{\partial \mathbf{x}}(\mathbf{x} - \mathbf{x}_0) + \frac{\partial \mathbf{f}}{\partial \mathbf{u}}(\mathbf{u} - \mathbf{u}_0), \quad (3)$$

or, by redefinition of the variables  $\mathbf{x} \rightarrow \mathbf{x} - \mathbf{x}_0 - \mathbf{f}(\mathbf{x}_0, \mathbf{u}_0) \times (t - t_0)$  and  $\mathbf{u} \rightarrow \mathbf{u} - \mathbf{u}_0$ ,

$$\dot{\mathbf{x}}(t) \approx \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \quad (4)$$

where  $\mathbf{A}$  is constant  $n \times n$  matrix of partial derivatives  $\partial f_i / \partial x_j$  and  $\mathbf{B}$  is constant  $n \times p$  matrix of partial derivatives  $\partial f_i / \partial u_j$ , and all partial derivatives are evaluated at the operating point.

For the linear system (4), there is a well-developed and extensive theory of the system behavior as a function of its structure, expressed in the matrices  $\mathbf{A}$  and  $\mathbf{B}$ . One may

broadly distinguish two parts of the theory, named classical control theory (e. g., [38]) and modern linear systems theory (e. g., [4,31]). We return to the classical control theory in the next section.

Modern control theory or linear systems theory (LST) is concerned with the dynamical properties of the system as a direct function of the system matrices  $\mathbf{A}$  and  $\mathbf{B}$ . A key element in this theory is the notion of the system *eigenvalues*, i. e., the eigenvalues of the matrix  $\mathbf{A}$ . If, for simplicity, we restrict ourselves to the endogenous dynamics of the system (set  $\mathbf{u} = 0$ ), we can write the solution to (4) as

$$\mathbf{x}_i(t) = c_{i,1} \exp(\lambda_1 t) + c_{i,2} \exp(\lambda_2 t) + \dots + c_{i,n} \exp(\lambda_n t), \quad i = 1, \dots, n, \quad (5)$$

where  $\lambda_1, \dots, \lambda_n$  are the  $n$  eigenvalues of the matrix  $\mathbf{A}$  and  $c_{i,j}$  are constants that depend upon the eigenvectors and the initial condition of the system. In other words, the resulting behavior is a weighted sum of distinct *behavior modes*,  $\exp(\lambda t)$ . If an eigenvalue is real, the corresponding behavior mode is exponential growth (if  $\lambda > 0$ ) or exponential decay (if  $\lambda < 0$ ). Complex-valued eigenvalues come in complex conjugate pairs  $\lambda = \tau \pm i\omega$  which give rise to oscillations  $\exp(\tau t) \sin(\omega t + \phi)$  of frequency  $\omega$  that are either expanding (if  $\tau > 0$ ) or damped (if  $\tau < 0$ ).

In this manner, the eigenvalues serve as a compact and rigorous characterization of the behavior (of linear systems).

At any point in time, any system, linear or nonlinear, may be approximated by the expression (5). Whether it remains a good approximation depends upon how much and how quickly the eigenvalues change due to the nonlinearities in the function  $f$ . If they are more or less constant for significant periods of time, we may speak of *quasilinear systems* that are well approximated by the linear system. In some cases, however, the eigenvalues change so rapidly that it makes little sense to characterize the behavior by equation (5). (See [29] for further discussion).

### Traditional Control Theory Approaches

The first set of methods, which we call the traditional approach, has been used for decades and is part of the standard curriculum in system dynamics teaching at the graduate level. It involves using the concepts from classical control theory [38] to very simple systems with only a few state variables.

The starting point is the simple first- and second-order positive and negative feedback loops found in any introductory treatment of system dynamics. The advantage of the approach is its simplicity. Although it serves at a guide to intuition, however, the obvious shortage is that it applies rigorously only to simple systems. There have been some attempts to treat higher-order systems by adding a few feedback loops [23], but the step to large-scale models is beyond this method given its inherent limitations.

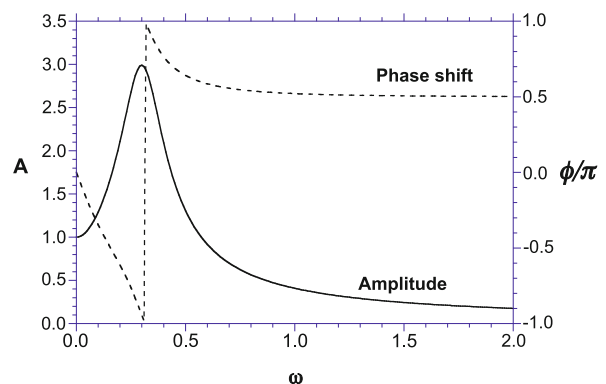
Graham [23] distills a number of principles that are based on the metaphor of a “disturbance” traveling along the chain of causal links in a feedback loop and getting amplified, damped, and possibly delayed in the process. For major negative feedback loops, which are known to tend to produce oscillation, adding minor negative loops and cross-links, or shortening the delay times increases the damping. Conversely, adding positive loops in to the oscillatory system tends to lengthen the period of oscillation whereas the effect on the damping depends upon the delays in the positive loop. Using the metaphor of pushing a child on a swing, it becomes clear that the timing of the propagation of a disturbance has as much importance for its effect on the damping as its strength.

For analyzing the behavior of positive feedback loops, Graham suggested calculating the Open-loop steady-state gain (OLSSG), a measure of the amplification around the loop. A gain greater than unity will result in exponential growth while gains less than 1 will give exponential adjustment (leveling off or decay). The intuition is perhaps best illustrated by an example: sales-driven growth. Suppose

a salesperson can eventually pull in \$100,000 per month in orders (probably with a several-month long delay), and assume that the company allocates 10% of revenue to marketing. Then this eventually leads to \$10,000 per month for sales efforts. If the cost of a salesperson (salary, overhead, expenses etc.) is, say, \$8,000 dollars per month, then the efforts of the current sales force will provide enough revenue to support  $10,000/8,000 = 1.25$  persons per current person. Thus, the OLSSG of the positive loop from salespersons  $\rightarrow$  orders  $\rightarrow$  revenues  $\rightarrow$  marketing budget  $\rightarrow$  salespersons is 1.25, and the system will grow exponentially (until other factors limit the growth). Conversely, if the gain is less than 1, one salesperson will not sell enough to support their own cost, and the loop will lead to exponential decay. Graham showed how the actual rate of growth is partly determined by the OLSSG, and partly by the time constants (delays etc.) involved. (See also Subsect. 15.3 in [59]).

In the context of oscillating systems, system dynamics has also employed a concept from classical control theory, frequency response. The frequency response is determined from the transfer function of the system,  $G(i\omega)$ , which is a complex-valued function that specifies how an input signal  $u(t)$  with frequency  $\omega$  results in an output signal  $x(t)$  that may be phase shifted (delayed), and either amplified or attenuated. For linear systems,  $G$  can be calculated directly from the system matrices in (4) – the transfer function (matrix) is  $G(i\omega) = B(i\omega I - A)^{-1}$ , where  $I$  is the identity matrix (see e. g. [4]). For nonlinear systems,  $G$  may be found through simulation experiments.

Usually,  $G$  is represented in a *Bode* or *phase-and-gain* diagram. For instance, Fig. 4 shows a Bode diagram of the



System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 4

Phase-and-gain diagram (Bode diagram) showing the inventory  $A \sin(\omega t + \phi)$  with amplitude  $A$  and phase shift  $\phi$ , relative to a sinusoidal demand  $\sin(\omega t)$ , for varying values of the frequency  $\omega$  of the demand fluctuation



inventory variable  $INV(t)$  relative to the exogenous demand input variable  $DEM(t)$  in the inventory workforce model in Fig. 3. The diagram shows how the relative amplitude of the oscillation and the relative phase shift (in radians) between input and output varies as a function of the frequency of the input.

It is clear from the diagram that there is a certain frequency range, around the system's own natural frequency, where fluctuations in demand are greatly amplified compared to other frequencies. Indeed, it is a general phenomenon in systems that they will tend to amplify certain frequencies while attenuating other frequencies. This may be used to explain or understand the role of particular structures in the model in generating oscillation at certain frequencies, even when there are no oscillations coming in from the outside world. (External random noise is enough to produce oscillations in the system because random noise contains fluctuations at all frequencies). In this manner, the approach nicely demonstrates the "endogenous viewpoint" that behavior (oscillations) is generated internally by the system. As an analytic tool for large scale systems, however, the method does not seem to produce any additional insights. Thus, we may conclude that the classical approaches serve mostly as intuitive metaphors to guide the analyst rather than as full analytical tools.

### Pathway Participation Metrics

The *pathway participation* method [34,35] represents a further development of an original suggestion by Richardson [46] to provide a rigorous definition of loop polarity and loop dominance. Richardson motivated this with the common confusion associated with positive feedback loops, which may exhibit a wide range of behaviors [23], as Barry Richmond noted with wonderful humor:

"Positive loops are ... er, well, they give rise to exponential growth ... or collapse ... but only under certain conditions ... Under other conditions they behave like negative feedback loops ..." [49].

Richardson proposed that the polarity of a loop be defined as the sign of the expression

$$\frac{\partial \dot{x}_i}{\partial x_i} = \frac{\partial f_i(x, u)}{\partial x_i}, \quad (6)$$

in the model (1), with a positive sign indicating a positive loop and vice-versa. When several loops operate simultaneously, the sign of the expression indicates whether the positive or negative loops dominate. Note, however, that the definition only applies to minor loops (i. e. loops involving a single level). Put differently, it only considers

the diagonal elements of the matrix  $A$  in the linearized system (4). Richardson [46] demonstrates how even with this limitation, analyzing the system with this metric can (sometimes) yield insights into behavior of higher-order systems.

The expression (6) hints that it is relevant to consider the curvature, i. e., the second time derivative,  $\ddot{x}$ , of a variable when looking for dominant structure. Although he does not say so explicitly, this is effectively the focus of Mojtahedzadeh's pathway method. Figure 5 shows how one may classify behavior by comparing the first and second time derivatives of a variable. As seen in the figure, the sign of the expression  $\ddot{x}/\dot{x}$ , which Mojtahedzadeh denotes the total *pathway participation metric* or *PPM*, indicates whether the behavior appears dominated by positive or negative loops, much in line with Richardson's definition of dominant polarity. A zero curvature indicates a shift in loop dominance (cf. the middle column in the figure). Note, however, that the interpretation of the middle row in the figure where the slope  $\dot{x}$  is zero has no clear interpretation in terms of loop dominance. Indeed this hints at one of the weaknesses of the approach that we will return to below.

Mojtahedzadeh's method proceeds by decomposing the PPM into its constituent terms as follows,

$$PPM_i = \frac{\ddot{x}_i}{\dot{x}_i} = \sum_{j=1}^n \frac{\partial f_i}{\partial x_j} \frac{\dot{x}_j}{\dot{x}_i}, \quad (7)$$

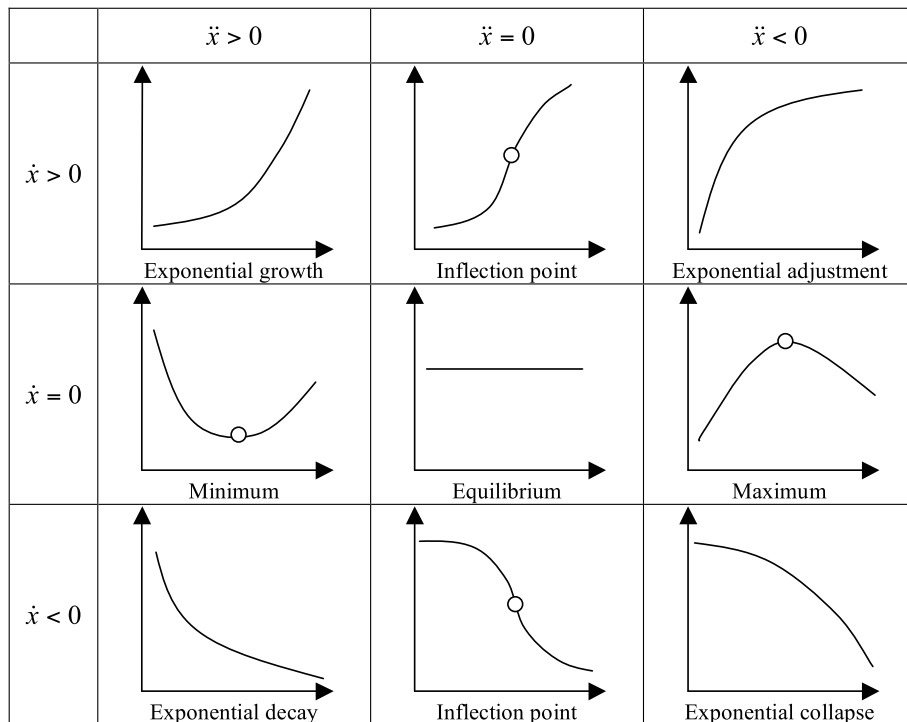
where, for brevity, we have chosen to ignore the exogenous variables  $u$ . One might say that each of the terms in the sum in (7) represents the separate influence of each of the systems' state variables on the behavior of  $x_i$ . Mojtahedzadeh in fact uses a normalized measure for the terms,

$$\frac{(\partial f_i / \partial x_j) \dot{x}_j}{\sum_{k=1}^n |(\partial f_i / \partial x_k) \dot{x}_k|}, \quad (8)$$

which can vary between  $-1$  and  $+1$ , to measure the relative importance of the pathway from variable  $j$ . By explicitly considering auxiliary variables  $y$  in the model, one may further decompose each term  $\partial f_i / \partial x_j$  into a sum of terms

$$\frac{\partial f_i^k}{\partial x_j} = \frac{\partial f_i}{\partial y_1} \cdot \frac{\partial y_1}{\partial y_2} \cdot \dots \cdot \frac{\partial y_{m-1}}{\partial y_m} \cdot \frac{\partial y_m}{\partial x_j}, \quad (9)$$

corresponding to a causal chain or pathway  $\pi_k = \{x_j \rightarrow y_m \rightarrow \dots y_2 \rightarrow y_1 \rightarrow \dot{x}_i\}$ . Mojtahedzadeh now considers each possible pathway (9) and defines the "dominant" pathway as the one with the largest numerical value



System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 5  
Characteristic behavior patterns based on the first and second time derivatives

and the same sign as  $PPM_i$ . Having selected this dominant pathway,  $\pi_{ij}^* = \{x_j \rightarrow y_m \rightarrow \dots y_2 \rightarrow y_1 \rightarrow \dot{x}_i\}$ , which originates in the state variable  $x_j$  the procedure is repeated for that state variable  $x_j$ , and so forth, until one either reaches one of the already “visited” state variables (in which case a loop has been found) or an exogenous variable (in which case an external driving force has been found). Thus, the procedure may result in three alternative forms of dominant structure illustrated in Fig. 6, namely a “pure” minor or major feedback loop, a pathway from a feedback loop elsewhere in the system, or a pathway from an exogenous variable.

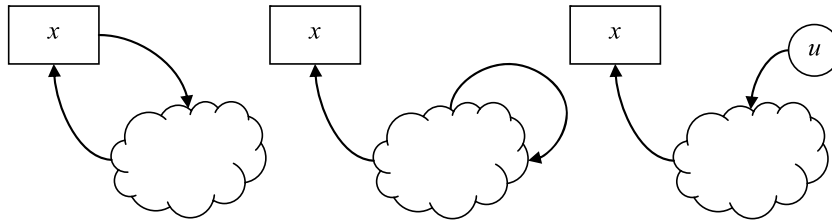
By dividing the observed model behavior into different phases according to the taxonomy in Fig. 5 and then applying the method just described at different points in during these phases, one can reveal how the dominant structure changes over time. For illustration, the PPM method is applied to the Bass model and the results are presented in Fig. 7. The figure shows the metrics of four alternative pathways (four feedback loops) and the results accord nicely with the informal analysis done earlier: The method identifies two phases, exponential growth, exponential adjustment, and identifies the “word-of-mouth” positive loop (loop 1) as dominant in the first phase and

the “exhaustion” loop (loop 2) as dominant in the second phase.

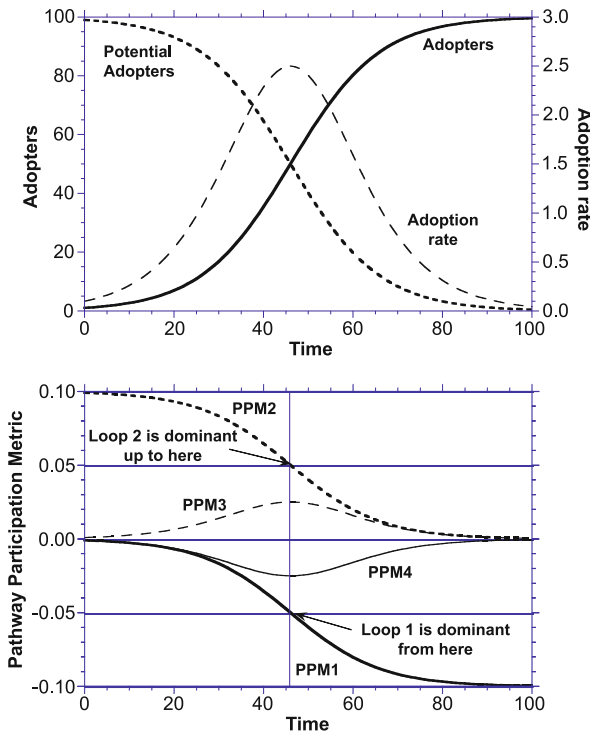
The PPM method is still mostly used at an early explorative stage on rather simple models, where it does appear to aid insight into the dynamics (e. g. [41]), and has been implemented in a software package, *Digest*, [35].

From the studies performed so far, it is clear that the main strength of this method is its relative computational simplicity (it does not require computing eigenvalues, which is a numerically demanding task), and the intuitive and direct connection it makes between the observed behavior and the influencing structural elements. Unlike the other approaches which operate in the “frequency domain”, the method considers the time path of a specific variable directly.

There are, however, some important outstanding issues that remain to be clarified. First, the method is not suitable for oscillatory systems. The problem is easy to recognize when one considers how the PPM measure will vary over the course of a sinusoidal outcome: The sign of the PPM will shift twice during each cycle, indicating that the behavior is alternately dominated by positive and negative loops, even though the system structure, and hence the loop dominance, may remain unchanged all the time.



System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 6  
Three alternative forms of dominant structure in the PPM method



System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 7  
Pathway participation measures in the Bass model

Richardson [46] already alluded to this problem by noting that the measure only considers the diagonal elements in the system matrix in (4), yet we know that the structure causing oscillation is the major negative loop that involve the off-diagonal elements. This is a significant limitation, given the prevalence and importance of oscillation in system dynamics analysis.

A second limitation of the current implementation of PPM is that it uses a depth-first search for the single most influential pathway for a variable. This strategy does not capture the situation where more than one structure may contribute significantly to the model behavior and,

through the depth-first algorithm, may miss alternative paths that could prove to yield a larger total value of the metric. This problem could be addressed by modifying the search algorithm and is most likely of minor importance.

Another issue is how to treat the case when  $\dot{x} = 0$  since it appears in the denominator of the terms in (6). However, it is not clear that it is necessary to do this division, given that it is easy to identify the nine cases in the figure by simply examining its sign. Thus, the issue is probably not of much significance.

The fourth issue, on the other hand, is more significant, namely the emphasis on identifying a single “dominant” structure. In reality, of course, the behavior of a variable is influenced by many loops and pathways at once. Reducing the consideration to a single one of these may miss important features of the structure-behavior relationships. For instance, a variable may be influenced by two negative loops and one positive, with the sum of the two negative loops dominating the influence of the positive loop, even though that loop by itself has the strongest influence on the behavior. It is more appropriate to consider the relative importance of alternative pathways, yet the method does not address how one would partition the behavior among pathways (the three structures in Fig. 6) – only among individual links.

Thus, while the notion of pathways seems an interesting and useful idea, it may be that it will ultimately be more effective to use a list, ranked in order of magnitude, of the pathways that influence a variable.

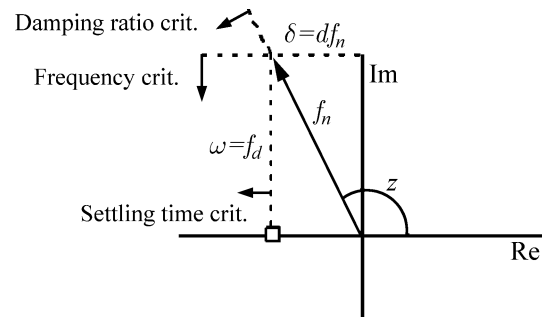
Finally, the method shares a weakness with the traditional method in that it considers primarily partial system structures rather than global system properties. In contrast, the two eigenvalue methods to which we now turn are based on a rigorous characterization of the entire system (at a given point in time).

### Eigenvalue Elasticity Analysis

The third method may be termed *eigenvalue elasticity analysis* (or EEA for short) and builds upon the tools

from modern linear systems theory (LST), applied to the linearized model (4). The method is concerned with the structural elements that significantly affect the system eigenvalues or behavior modes – the values  $\lambda$  in (5). Specifically, it measures influence by the elasticity of an eigenvalue  $\lambda$  with respect to some parameter  $g$  in the model, defined as  $\varepsilon = (\partial\lambda/\partial g)(g/\lambda)$ , i.e. the fractional change in the eigenvalue relative to the fractional change in the parameter. The advantage of this fractional measure is that it is dimensionless, i.e., independent upon the choice of units, including the time scale unit. Sometimes, the influence measure is used instead, defined as  $\mu = (\partial\lambda/\partial g)g$ . This measure has dimension [1/time] and so depends upon the choice of its time unit, but it is generally easier to interpret for complex-valued eigenvalues and avoids numerical problems with very small or zero eigenvalues (see [29,54]).

The idea behind EEA was first introduced in system dynamics by Forrester [14] in the context of economic stabilization policy. For purposes of policy analysis in oscillating systems, one may define a number of criteria from engineering control theory, all of which relate to the eigenvalues of the system, as summarized in Table 2. Figure 8 provides a graphical characterization of the eigenvalues and policy criteria in the complex plane. Though these measures are not new, the EEA method is unique in its attempt to use them to gain qualitative intuitive understanding of the system. A significant step in this direction was first suggested by Forrester [15] with the notion that the elasticities of any links in the model (corresponding to elements of the matrix  $A$  in the linearized system (4)), can be interpreted as the sum of elasticities of all feedback



**System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 8**

**Characterization of eigenvalues plotted in the complex plane**

loops containing that link. We have chosen to name this approach *loop eigenvalue elasticity analysis (LEEA)*.

Kampmann [28] provided a rigorous definition of LEEA and also pointed to the fact that feedback loops are not independent. In other words, given the possibly very large number of loops in a given model (Kampmann demonstrated how the theoretical maximum number of loops grows combinatorically with the number of variables), it only makes sense to speak of individual contributions of a limited set of *independent* loops. He proved that a fully connected system (where there is a feedback loop between any pair of variables – the typical case in system dynamics models) with  $N$  links and  $n$  variables has a total of  $N - n + 1$  independent loops and provided a procedure for constructing this set and calculating the loop elasticities.

Kampmann's analysis points to a fundamental issue relating to the notion of feedback loops as a way to explain

**System Dynamics, Analytical Methods for Structural Dominance Analysis in, Table 2**

**Stabilization policy criteria and corresponding effects on eigenvalues and BDW of a policy change in a system element  $g$**

| Policy Criterion        | Description                                                                                                                                | Change in eigenvalue<br>$\lambda = \delta \pm i\omega, \omega > 0$ | Change in BDW $w$                               | Appropriate measure in time path |
|-------------------------|--------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------|-------------------------------------------------|----------------------------------|
| Damping                 | Increases the rate of decay of oscillation (or decreases the rate of expansion)                                                            | $\frac{\partial \delta}{\partial g} \frac{g}{\delta} < 0$          | N/A                                             | $\frac{x(t+T)}{x(t)}$            |
| Frequency               | Decreases the frequency of oscillation (or lengthens the period $T$ )                                                                      | $\frac{\partial \omega}{\partial g} \frac{g}{\omega} < 0$          | N/A                                             | $T$                              |
| Variance                | Reduces the variance of a target variable (or the weighted average variances of several variables)                                         | No simple relation                                                 | $\frac{\partial w}{\partial g} \frac{g}{w} < 0$ | $\int x(t)^2 dt$                 |
| Auto-spectrum           | Reduces variance of target variable(s) within a target frequency range                                                                     | No simple relation                                                 | $\frac{\partial w}{\partial g} \frac{g}{w} < 0$ | Filter in frequency domain       |
| Frequency response gain | Reduces the gain (amplification) in the target frequency range for a particular combination of disturbance exogenous and output variables. | Based upon transfer function $G(i\omega)$                          |                                                 |                                  |

**System Dynamics, Analytical Methods for Structural Dominance Analysis in, Table 3**

Loops and their influences in the inventory workforce model. Values are measured at time  $t = 0$ . The model contains three eigenvalues,  $\lambda_1 = -0.250$  and  $\lambda_2, \lambda_3 = -0.138 \pm i0.285$ . The influence measure is defined as  $g \cdot \partial \lambda / \partial g$ . For the imaginary part, a positive influence measure means that the frequency is increased

| Loop | Nodes                                           | Gain   | Influence on $\text{Re}[\lambda_1]$ | Influence on $\text{Re}[\lambda_2]$ | Influence on $\text{Im}[\lambda_2]$ |
|------|-------------------------------------------------|--------|-------------------------------------|-------------------------------------|-------------------------------------|
| 1    | ED > CED                                        | -0.250 | -0.250                              | 0.000                               | 0.000                               |
| 2    | W > HFR                                         | -0.200 | 0.000                               | -0.100                              | -0.022                              |
| 3    | INV > IC > DP > SP > EO > P                     | -0.076 | 0.000                               | -0.038                              | 0.008                               |
| 4    | INV > IC > DP > DW > HFR > W > NP > P           | -0.100 | 0.000                               | 0.000                               | 0.176                               |
| 5    | INV > IC > DP > DW > HFR > W > NP > SP > EO > P | 0.015  | 0.000                               | 0.000                               | -0.027                              |

behavior: the significance assigned to a particular loop depends upon the context (the chosen independent loop set). In other words, feedback loops are derived and relative concepts rather than fundamental independent building blocks of systems. Oliva [40] further refined the definition of independent loop sets by introducing the *Shortest independent loop set (SILS)* along with a procedure for constructing the set. Although a SILS is not generally unique, experience seems to suggest that it is easier to interpret [41]. Yet the issue remains that independent feedback loop sets are relative concepts.

In Table 3, we show how the LEEA analysis applies to the simple inventory-workforce model in Fig. 3. The model contains a total of 5 feedback loops, all of which are independent. The loops are listed in Table 3, including their constituent variables (nodes), and the gain of the loop (defined in a similar manner to the pathway participation metrics above). We see that there are three minor negative loops, related to the exponential smoothing of expected demand (loop 1) and the adjustment of workforce to desired workforce (loop 2). The minor loop 3 is the “overtime shortcut” that allows production to adjust part way to desired production immediately so one does not have to wait for the workforce to adjust. Loop 4 is the main major negative loop that adjusts inventory to desired levels via workforce adjustment. Finally, loop 5 (the only positive loop) is a fairly weak loop that moderates the effect of loop 4 by adjusting the overtime effect “back to normal” when the workforce is brought in line with desired production.

Although the model is nonlinear (due to the overtime function), the eigenvalues do not change very much over the course of its behavior. The model contains one real eigenvalue ( $\lambda_1 = -0.250$ ) and one pair of complex conjugate eigenvalues ( $\lambda_2, \lambda_3 = -0.138 \pm i0.285$ ). The first eigenvalue corresponds to the adjustment of expected demand (ED). The other pair produces a damped oscillation in inventory and workforce.

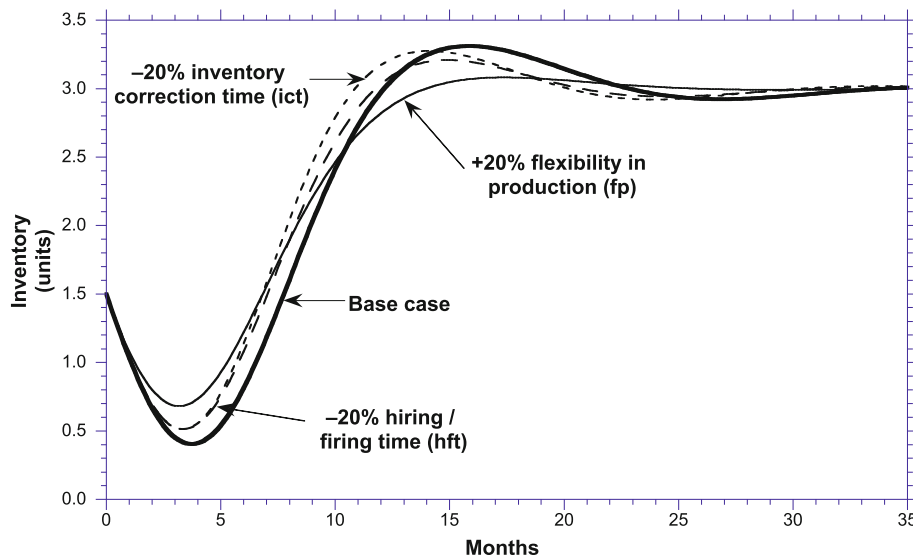
Table 3 also shows the loop influences upon the three eigenvalues. Note how there is a one-to-one correspondence between loop 1 (the adjustment of expected demand) and the first eigenvalue. This is due to fact that the ED level constitutes a single strongly connected component of the model (see Fig. 3), i.e. there is no feedback between this level and the rest of the model. We also note that the workforce adjustment and the overtime loops have a stabilizing influence upon the behavior (they make the real part of the oscillatory eigenvalues more negative and have relative little effect upon the frequency of oscillation). Conversely, the major negative loop 4 has a destabilizing influence, since strengthening it will increase the frequency of oscillation and not increase the damping. The effects of loop 5 are fairly weak.

From this analysis, one would therefore expect parameters that strengthen loop 2 (shortening hire/fire time) or loop 3 (increase overtime effect) would stabilize the system while strengthening loop 4 (shorter inventory adjustment time) will destabilize the system. Indeed this is what happens, as illustrated in the simulations in Fig. 9.

The EEA/LEEA method has been applied in a number of contexts (e.g. [1,20,22,24,29,51,52,54]), but remains a tool employed only by specialists in fundamental research, not least because it has not been incorporated into standard software packages. Thus, the potential of the method for widespread practice remains unexplored.

One might be skeptical that a method derived from linear systems theory may have any use for the nonlinear models found in system dynamics. Kampmann and Oliva [29] considered what types of models the method would be particularly suited for. They defined three categories of models, based upon the behavior they are designed to exhibit: 1) linear and quasilinear models, 2) nonlinear single-transient models, and 3) nonlinear periodic models. The first category encompasses models of oscillations, possibly combined with growth trends, with relatively stable equilibrium points, (e.g., the classical in-





System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 9

Simulated behavior of inventory-workforce model, showing the effect on inventory of parameter changes for overtime (flexibility of production), inventory adjustment time (ict), and labor hiring/firing time (hft), respectively

dustrial dynamics models [11]). Nonlinearities may modify behavior (particularly responses to extreme shocks) but the instabilities and growth trends can be analyzed in terms of linear relationships. Kampmann and Oliva concluded that LEEA showed the most promise and potential for this class of models because the analytical foundations are solid and valid, and because the method has the ability to find high-elasticity loops even in large models very quickly without much intervention on the part of the analyst.

The second class is typical of scenario models like the World Model [13,33], the Urban Dynamics Model [12], or the energy transition model in [57], to name a few, that show a single transient behavior pattern, like overshoot and collapse or a turbulent transition to a new equilibrium. In these models, nonlinearities usually play an essential role in the dynamics. Yet it is possible to divide the behavior into distinct phases where certain loops tend to dominate the behavior. In this class of models LEEA also shows promise by measuring shifts in structural dominance by the change in elasticities. But it requires more input from the analyst (e. g. in defining the different phases of the transition) and it has no obvious advantage over other methods, like PPM.

The third class, nonlinear periodic models, are those that exhibit fluctuating behavior in which nonlinearities play an essential role, such as like limit cycles, quasiperiodic behavior, or chaos, (see, e. g. [48]). Here the utility of the method is much less clear and depends upon the

specifics of the model in question. For example, the classic Lorenz model that exhibits limit cycles, period doubling and deterministic chaos does not lend itself to any insight using LEEA [29]. This is particularly the case in systems with strong nonlinearities such as min and max functions. In these systems, the behavior may change abruptly (eigenvalues suddenly shift) in what is called border-collision bifurcations [37,61]. In other cases, the method of breaking the behavior into phases with different dominant structures may yield significant insight from LEEA. For instance, Sterman's simple long wave model [58] lends itself well to this approach (e. g. [24,28]).

In the present paper, we add a fourth category of models or behavior for which the method has not been explored yet. We name this category nonlinear multi-modal models. These encompass the cases where one behavior mode interacts with and therefore modifies another behavior mode – something that can only happen in nonlinear systems. The most common example is mode-locking or entrainment, in which oscillations become synchronized (e. g. [25]). Another example is mode modification, where one behavior mode (growth or oscillation) affects the character of another (typically oscillation). An example of this is the interaction of the business cycle with the economic long wave, where the former tends to get more severe during long wave downturns [16]. Whether LEEA can contribute to this class of models remains to be seen.

Compared to the former two methods, the EEA/LEEA is mathematically more general and rigorous, though

many of the mathematical issues in the method remain to be addressed, as we summarize below. This rigor is also the main strength of the method, since it provides an unambiguous and complete measure of the influence of the entire feedback structure on all behavior modes.

A weakness or challenge that is starting to show up is the computational intensity in calculating eigenvalues and elasticities. This is not so much an issue of computer time and memory space as of the stability of numerical methods. Kampmann and Oliva [29] found that the numerical method used sometimes proved unstable, yielding meaningless results. Clearly, there is a need to explore this issue further, possibly building upon the developments in control engineering.

A more serious weakness is the difficulty in interpreting the results: Eigenvalues do not directly relate to the observed behavior of a particular variable. The concepts of eigenvalues and elasticities are rather abstract and unintuitive [10]. There is a need for tools and methods that can translate them into visible, visceral, and salient measures. Here, the measures in Table 2 may provide a guide. In particular, it is possible to use (linear) filtering in the frequency domain to define a behavior of interest. For example, an analyst may be concerned with structures causing a typical business cycle (3–4-year oscillation) and, by specifying a filter that “picks out” that range of fluctuation, could obtain measures for structures that have elasticities in that range. Because filters are typically linear operators, all the analytical machinery of the LEEA method will also apply in this case – a significant advantage.

Using filters will also solve an issue that appears in large-scale models, namely the presence of several identical or nearly identical behavior modes. Saleh et al. [54] do consider the analytical problems associated with repeated eigenvalues, where it becomes necessary to use generalized eigenvectors, and where other behavior modes appear involving power functions of time. A filter essentially constitutes a weighted average of behavior modes and in this fashion avoids the “identity problem” of non-distinct eigenvalues.

The most serious theoretical issue, in our view, is how the results are interpreted using the feedback loop concept. As mentioned, the concept is relative (to a choice of an independent loop set). Moreover, practice reveals that the number of loops to consider is rather large and that the loops elasticities often do not have an easy or intuitive explanation. A lot of care must be taken when interpreting the results. For instance, Kampmann and Oliva [29] found that “phantom loops” – loops that cancel each other by logical necessity and are essentially artifacts of the equation formulations used in the model – could nonetheless

have large elasticities and thus seriously distort the interpretation of the results. An example of “phantom loops” is found in the Bass model in Fig. 1, where loops 3 and 4 are artifacts of the way the model is formulated. If the variable *Total population* (T) was eliminated from the equations, the loops would disappear and in fact they exactly cancel each other out (since T is constant). Nonetheless, they appear on the list of loops and appear to have a separate influence on behavior. These kinds of problems may not be intractable, but their resolution will require careful mathematical analysis.

Finally, a problem with EEA and LEEA is that it only considers changes to behavior modes, not the degree to which these modes are expressed in a system variable of interest. This issue is addressed by also considering the *eigenvectors* of the system, which is the foundation for the analysis in the next section.

### Eigenvectors and Dynamic Decomposition Weights (DDW)

The last set of methods, which are still in early development, we have termed the *eigenvector-based* approach (EVA). EVA attempts to improve the EEA/LEEA method by considering how much an eigenvalue or behavior mode is expressed in a particular system variable. The logic of the method and how EEA and EVA complement each other is shown in Fig. 10. As shown by Kampmann [28], in a sense there is a one-to-one correspondence between eigenvalues and loop gains whereas the eigenvectors arise from the remaining “degrees of freedom” in the system. The observed behavior of the state variables in the model is then the combined outcome of the behavior modes (from the loop gains) and the weights for each mode (from the eigenvectors) in the respective state variable.

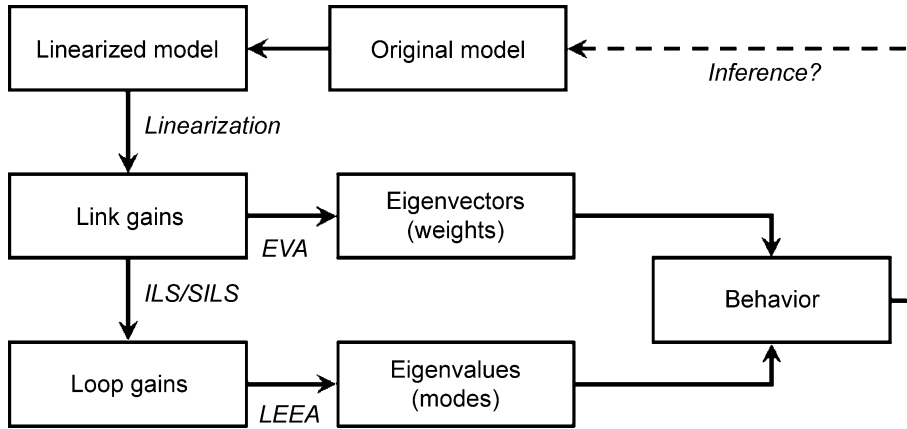
A number of researchers have attempted to develop EVA methods. Some emphasize the curvature (second time derivative) of the behavior, similar to the starting point of the PPM method [24,50,51,52]. The slope or rate of change  $\dot{x}(t)$  of a given variable  $x$  in the linearized system may be written by

$$\dot{x}(t - t_0) = w_1 \exp(\lambda_1(t - t_0)) + \dots + w_n \exp(\lambda_n(t - t_0)) \quad (10)$$

where the weights  $w_i$  are related to the eigenvectors. Then the curvature at time  $t_0$  is

$$\ddot{x}(t_0) = w_1 \lambda_1 + \dots + w_n \lambda_n \quad (11)$$

One may therefore interpret (11) has the sum of contribution from individual behavior modes. Güneralp [24]



**System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 10**  
Schematic view of eigenvalue and eigenvector analysis approach

suggested using the terms on the right-hand side of (11) as weights to combine elasticities of individual behavior modes  $\varepsilon_i$  with respect to some system element (like a link gain or a loop gain) into a weighted sum

$$\bar{\varepsilon} = \frac{\sum_{i=1}^n w_i \lambda_i \varepsilon_i}{\sum_{i=1}^n |w_i \lambda_i|}, \quad (12)$$

as a measure of the overall significance of that system element. He further normalized the elasticity measure by the elasticity measure for other system elements, i. e., assuming there are  $K$  such elements (loops or links), the relative importance  $\rho_k$  of the  $k$ th element is defined as

$$\rho_k = \frac{\bar{\varepsilon}_k}{\sum_{j=1}^K |\bar{\varepsilon}_j|}, \quad (13)$$

with the motivation that elasticities may vary greatly in numerical values, making comparisons at different points in time difficult, whereas  $\rho_k$  is a relative measure varying between +1 and -1. His results shed an alternative light on the behavior of these models, but the mathematical meaning, consistency and significance of the doubly normalized measure (13) remains to be clarified. It is still too early to tell what the most useful approach will be, but one may note that the emphasis on the curvature shares the basic weakness in the PPM approach in dealing with oscillations.

Other researchers have looked directly at the *dynamic decomposition weights* (DDW)  $w_i$  in (10), i. e., the relative weight of the modes for a particular variable, from a policy

criterion perspective, similar to Forrester's original focus and the starting point for the EEA analysis [21,53,54].

For instance, Saleh et al. [54] look at how alternative stabilization policies affect the behavior of business cycle models, using both a simple inventory-workforce model [59], and a more extensive model based on Mass [32] and used in the LEAA analysis of Kampmann and Oliva [29]. Using the procedure in Fig. 9, they decompose the net stabilizing effect of a policy into its effect on the behavior mode itself (LEEA) and its effect on the expression of that mode in the variable of interest, measured the dynamic decomposition weights (EVA or DDW).

To illustrate the approach we perform the computations for the inventory-workforce model (Fig. 3 and Table 1). We find that the following equations describe the behavior of the state variables

$$\begin{aligned} ED &= 1 - 0.500e^{-0.250t} \\ INV &= 3 - 2.167e^{-0.250t} \\ &\quad + 1.134e^{-0.138t} \sin(2.945 + 0.285t) \\ WF &= 1 + 0.669e^{-0.250t} \\ &\quad - 1.169e^{-0.138t} \sin(1.553 - 0.285t). \end{aligned} \quad (14)$$

As expected from the structure of the model, the behavior of *Expected Demand* does not have an oscillatory component and only shows a short transient exponential adjustment for the stock to match *Demand*. On the other hand, *Inventory* and *Workforce*, in addition to having the transient behavior to reach equilibrium captured by the first eigenvalue, have an oscillatory component represented by the second eigenvalue. Note that each state variable has a different *Dynamic Decomposition Weight* ( $w$ ) for each

## System Dynamics, Analytical Methods for Structural Dominance Analysis in, Table 4

Elasticity to parameters of weight of eigenvalue 2 ( $-0.138 + i0.85$ ) on inventory and influence of parameters on eigenvalue 2 – inventory–workforce model

| Parameter                   | $w_2$ on INV Elasticity | Influence on $\text{Re}[\lambda_2]$ | Influence on $\text{Im}[\lambda_2]$ |
|-----------------------------|-------------------------|-------------------------------------|-------------------------------------|
| Demand                      | 2.000                   | 0.000                               | 0.000                               |
| Inventory correction time   | 0.656                   | 0.038                               | -0.157                              |
| Flexibility in production   | 0.364                   | -0.549                              | -0.266                              |
| Productivity                | -0.353                  | 0.000                               | 0.000                               |
| Time to change expectations | -0.240                  | 0.000                               | 0.000                               |
| Normal inventory coverage   | 0.239                   | 0.000                               | 0.000                               |
| Hiring/Firing time          | 0.238                   | 0.100                               | -0.127                              |
| Disequilibrium fraction     | 0.000                   | 0.000                               | 0.000                               |

reference mode, i. e., each eigenvalue contributes differently to the overall behavior of each state variable.

An exploration of the policy design space can be achieved by assessing the influence of model parameters on the dynamic decomposition weight. By focusing on the weights of the behavior modes for the variable of interest we can identify leverage points to increase or decrease the presence of a behavior mode in the variable. The weight elasticity column in Table 4 reports the parameter elasticity of  $w_2$  (the weight of eigenvalue 2, the oscillatory behavior mode) on *Inventory* ( $\varepsilon_w = (dw/dp)(p/w)$ ). The magnitude of the elasticity quantifies the impact that changes in the parameter value have on the weight of the oscillatory behavior model on *Inventory*. The table is sorted in descending order of absolute value of elasticity.

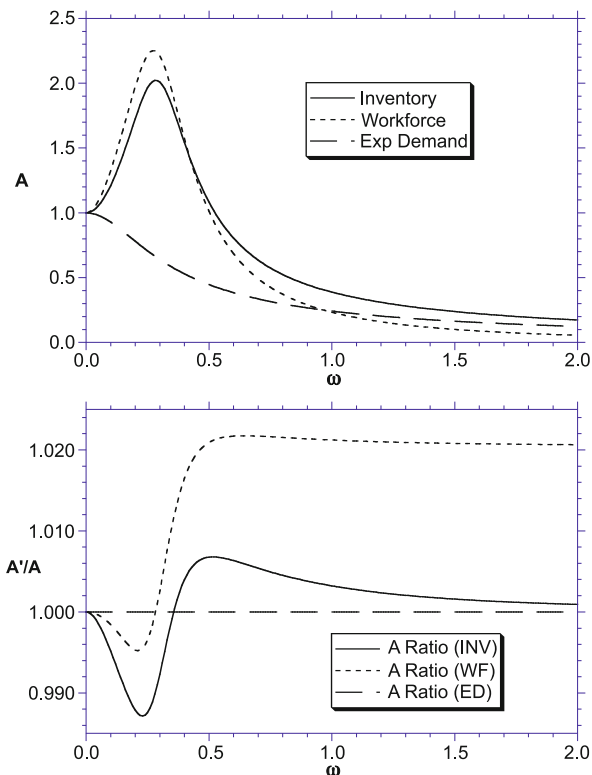
Changes in parameters, however, not only impact the behavior decomposition weights, but also change the eigenvalues themselves. This dual impact of parameter changes introduces a challenge in developing policy recommendations. The last two columns of Table 4 report the influence on the eigenvalue (real and imaginary part) for each parameter. These measures of influence should be interpreted in a similar way as the weight elasticities. The influence measure is defined as  $\mu_\lambda = (\partial\lambda/\partial p)p$ . A positive real-part measure indicates that increasing the parameter will destabilize the system by lengthening the settling time and vice-versa. A positive imaginary-part measure indicates that increasing the parameter will increase the frequency of oscillation – normally considered a destabilizing influence – and vice-versa.

Five parameters, *demand*, *disequilibrium fraction*, *productivity*, *normal inventory coverage*, and *time to change expectations*, have no influence on the oscillatory behavior mode. *Demand* and *disequilibrium fraction* are initialization constants that do not participate in any of the feedback loops in the model. *Productivity* is essentially a scaling measure having to do with the definition of units of

labor and goods in the model. Redefining units should not affect the dynamics of the model. While *time to change expectations* is involved in loop 1, it does not participate in the oscillatory behavior observed in the model since, as discussed above, *Expected Demand* is in a separate strongly connected component of the model.

In accordance with LEEA, the *flexibility in production* parameter, which strengthens overtime loop 3 (cf. Fig. 3), has a strong stabilizing influence, by both increasing the damping and lowering the frequency of the oscillatory mode. Likewise, as predicted by LEEA, a shorter *hiring/firing time* will increase damping by strengthening the labor adjustment loop 5 but, again in accordance with LEEA, also increases the frequency of adjustment because it also strengthens the major loop 4. Finally, lowering the *inventory correction time* will strengthen the link from *inventory* to *desired production*, and consequently the three loops 3, 4 and 5, with the net effect that although the adjustment is a little faster (a more negative real part), the frequency is also increased significantly, i. e., it is a less effective way of stabilizing the system (cf. Fig. 9).

As an alternative approach, Fig. 11 shows what happens to the frequency response of the state variables (*Inventory* INV, *Workforce* WF, and *Expected Demand* ED) when the parameter Hiring/Firing Time (hft) is reduced by 2% from 5 to 4.9. There are a number of things to notice in the figure. First, there is no effect whatsoever on the ED variable, which should not be surprising, given that there is no feedback to this variable from the rest of the system. Second, the effect on the amplitude, like the amplitude itself, is strongly dependent upon the frequency of variation. We see that there is a significant amount of dampening on the *Inventory* fluctuation around the resonant frequencies in the range 0.1 to 0.3. On the other hand, there is a small amount of amplification of inventory in the higher frequency ranges. The effect on *Workforce* is very different: though there is a small attenuation in the reso-



**System Dynamics, Analytical Methods for Structural Dominance Analysis in, Figure 11**

Effect on frequency response of the inventory workforce model of reducing the parameter Hiring/Firing Time (hft) from 5.0 to 4.9. The diagram shows the gain of the base case (upper graph) for the three state variables, and the resulting change in the gain, measured as the ratio ( $A'/A$ ) from the parameter change (lower graph)

nant frequencies, there is a significant increase in variance in the higher frequency range. In other words, although the LEEA analysis showed a faster hiring policy to be stabilizing (by strengthening loop 2, cf. Table 3), the DDW analysis shows that it depends – both upon the variable in question and the context (frequency of variation).

### Future Directions

As mentioned above, it is not possible to construct a complete theory that will automatically provide modelers with “the” dominant structure. Given the analytical intractability of nonlinear high-order systems found in our field, the most we can hope for is a set of tools that will guide the analysis and aid the development of the modeler’s intuition.

That said, however, we are left with an impression that the analytical foundation for these tools is in need of further development before one rushes into implement-

ing them into software packages. We are quite satisfied with the current state of affairs in this regard, where code, models, and documentation are made freely to download (most of the cited papers provide a URL to their code and models). Understanding *how* and *why* the tools work the way they do is crucial, and this will require that a number of puzzles, uncertainties, and technical problems be addressed. Only then will the time come to submit the methods for wider application to test their real-world utility.

While the classical method remains a useful intuitive guide and teaching tool for graduate students, there are no signs that it may be developed further. (That said, it is possible that the classical control transfer function method may be employed in the eigensystem approaches to explore nested canonical systems, though this is purely speculative). The pathway method would benefit from a firmer mathematical foundation. In particular, it would be important to compare how its results and conclusions compare to those found in the LST. It is possible that the pathway method may eventually be merged with the LST approaches as a subset of a general analytical toolbox. We believe that there is a great deal of promise in combining the eigenvalue and eigenvector analysis in the LST approaches. This combination will yield a complete system characterization and an understanding of both how particular feedback loops are involved in generating a behavior mode, and how system elements determine the expression of that behavior mode in a particular variable. A unified LST approach along the lines suggested in Fig. 10 thus seems within reach.

It will probably be a while, however, before these methods will find their way into widely available and use-friendly software packages. Apart from the theoretical issues alluded to above, a number of technical issues related to numerical calculations, various “pathological cases” (such as non-distinct eigenvalues), and special cases of feedback loops (“figure-eight” loops, for instance), will need to be addressed.

On the more creative side, it would be interesting to explore alternative forms of visualizing the various influence measures developed. For instance, one could imagine that links between variables in a model diagram “glow” in different colors and intensities depending upon their effect on a behavior pattern in question. This is not just a question of fancy user interfaces: as mentioned in the introduction, the function of these tools will be as intuitive consistent aids to understanding, not analytical “answering machines”. In this light, the visualization is as important as the analytical principles behind it. Given the power of the human eye in finding patterns in visual data, this could be a significant next step.



## Bibliography

- Abdel-Gawad A, Abdel-Aleem B, Saleh M, Davidsen P (2005) Identifying dominant behavior patterns, links and loops: Automated eigenvalue analysis of system dynamics models. Proceedings of the Int System Dynamics Conference, Boston, July 2005. System Dynamics Society, Albany
- Barlas Y (1989) Multiple Tests for Validation of System Dynamics Type of Simulation Models. *Eur J Oper Res* 42(1):59–87
- Bass FM (1969) A new product growth for model consumer durables. *Manag Sci* 15(5):215–227
- Chen CT (1970) Introduction to Linear System Theory. Holt, Rinnehart and Winston, New York
- Eberlein R (1984) Simplifying Models by Retaining Selected Behavior Modes. Ph D Thesis, Sloan School of Management, MIT, Cambridge
- Eberlein RL (1986) Full Feedback Parameter Estimation. In: Proceedings of the Int Systems Dynamics Conference, Sevilla, Spain. System Dynamics Society, Albany, pp 69–83
- Eberlein RL (1989) Simplification and understanding of models. *Syst Dyn Rev* 5(1):51–68
- Eberlein RL, Wang Q (1985) Statistical Estimation and System Dynamics Models. Proceedings of the Int Systems Dynamics Conference, Keystone, USA. System Dynamics Society, Albany, pp 206–222
- Ford A, Flynn H (2005) Statistical screening of system dynamics models. *Syst Dyn Rev* 21(4):273–303
- Ford DN (1999) A Behavioral Approach to Feedback Loop Dominance Analysis. *Syst Dyn Rev* 15(1):3–36
- Forrester JW (1961) Industrial Dynamics. Productivity Press, Cambridge
- Forrester JW (1969) Urban Dynamics. Productivity Press, Cambridge
- Forrester JW (1971) World Dynamics. Productivity Press, Cambridge
- Forrester N (1982) A Dynamic Synthesis of Basic Macroeconomic Policy: Implications for Stabilization Policy Analysis. Ph D Thesis, Sloan School of Management, MIT, Cambridge
- Forrester N (1983) Eigenvalue analysis of dominant feedback loops. Proceedings of the Int System Dynamics Conference, Chestnut Hill, USA. System Dynamics Society, Albany
- Forrester JW (1993) System Dynamics and the Lessons of 35 Years. In: DeGreene KB (ed) Systems-Based Approach to Policymaking. Kluwer, Norwell, pp 199–240
- Forrester JW, Senge PM (1980) Tests for Building Confidence in System Dynamics Models. *TIMS Stud Manag Sci* 14:209–228
- Forrester JW, Mass NJ, Ryan CJ (1976) The System Dynamics National Model: Understanding Socio-Economic Behavior and Policy Alternatives. *Technol Forecast Soc Chang* 9(1–2):51–68
- Goldberg DE (1989) Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley, Reading
- Gonçalves P (2003) Demand bubbles and phantom orders in supply chains. Ph D Thesis, Sloan School of Management, MIT, Cambridge
- Gonçalves P (2008) Behavior modes, pathways and overall trajectories: Eigenvalue and eigenvector analysis in system dynamics. *Syst Dyn Rev*, forthcoming
- Gonçalves P, Lerpattarapong C, Hines JH (2000) Implementing formal model analysis. Proceedings of the Int System Dynamics Conference, Bergen, Norway, August 2000. System Dynamics Society, Albany
- Graham AK (1977) Principles on the Relationship Between Structure and Behavior of Dynamic Systems. Ph D Thesis, Sloan School of Management, MIT, Cambridge
- Güneralp B (2006) Towards Coherent Loop Dominance Analysis: Progress in Eigenvalue Elasticity Analysis. *Syst Dyn Rev* 22(3):263–289
- Haxholdt C, Kampmann CE, Mosekilde E, Sterman JD (1995) Mode Locking and Entrainment of Endogenous Economic Cycles. *Syst Dyn Rev* 11(3):177–198
- Holland JH (1992) Adaptation in Natural and Artificial Systems. MIT Press, Cambridge
- Homer JB (1983) Partial-Model Testing as a Validation Tool for System Dynamics. In: Proceedings of the Int System Dynamics Conference, Chestnut Hill, USA, July 1983. System Dynamics Society, Albany, pp 920–932
- Kampmann CE (1996) Feedback Loop Gains and System Behavior (unpublished manuscript). In: Proceedings of the Int System Dynamics Conference, Cambridge, USA, July 1996. System Dynamics Society, Albany, pp 260–263
- Kampmann CE, Oliva R (2006) Loop Eigenvalue Elasticity Analysis: Three Case Studies. *Syst Dyn Rev* 22(2):146–162
- Lane DC, Smart C (1996) Reinterpreting 'generic structure': evolution, application and limitations of a concept. *Syst Dyn Rev* 12(2):87–120
- Luenberger DG (1979) Introduction to Dynamic Systems: Theory, Models and Applications. Wiley, New York
- Mass NJ (1975) Economic Cycles: An Analysis of Underlying Causes. Productivity Press, Cambridge
- Meadows DH, Meadows DL, Randers J, Behrens III WW (1972) The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind. Universe Books, New York
- Mojtahedzadeh MT (1996) A path taken: Computer-assisted heuristics for understanding dynamic systems. Ph D Thesis, Rockefeller College of Public Affairs and Policy, State University of New York at Albany, Albany
- Mojtahedzadeh MT, Andersen D, Richardson GP (2004) Using Digest to implement the pathway participation method for detecting influential system structure. *Syst Dyn Rev* 20(1):1–20
- Morecroft JDW (1985) Rationality in the Analysis of Behavioral Simulation Models. *Manag Sci* 31(7):900–916
- Mosekilde E, Laugesen JL (2006) Nonlinear Dynamic Phenomena in the BEER Model. Department of Physics, The Technical University of Denmark, Kongens Lyngby
- Ogata K (1990) Modern Control Engineering, 2nd edn. Prentice Hall, Englewood Cliffs
- Oliva R (2003) Model Calibration as a Testing Strategy for System Dynamics Models. *Eur J Operat Res* 151(3):552–568
- Oliva R (2004) Model Structure Analysis Through Graph Theory: Partition Heuristics and Feedback Structure Decomposition. *Syst Dyn Rev* 20(4):313–336
- Oliva R, Mojtahedzadeh M (2004) Keep it simple: Dominance assessment of short feedback loops. Proceedings of the Int System Dynamics Conference, Oxford, UK, July 2004. System Dynamics Society, Albany
- Ott E (1993) Chaos in Dynamical Systems. Cambridge University Press, New York

43. Peterson DW (1980) Statistical Tools for System Dynamics. In: Randers J (ed) *Elements of the System Dynamics Method*. Productivity Press, Cambridge, pp 224–241
44. Peterson DW, Eberlein RL (1994) Reality Checks: A Bridge Between Systems Thinking and System Dynamics. *Syst Dyn Rev* 10(2/3):159–174
45. Radzicki MJ (2004) Expectation Formation and Parameter Estimation in Uncertain Dynamical Systems: The System Dynamics Approach to Post Keynesian-Institutional Economics. *Proceedings of the Int System Dynamics Conference*, Oxford, UK, July 2004. System Dynamics Society, Albany
46. Richardson GP (1984/1995) Loop Polarity, Loop Dominance, and the Concept of Dominant Polarity. *Syst Dyn Rev* 11(1):67–88
47. Richardson GP (1986) Dominant structure. *Syst Dyn Rev* 2(1):68–75
48. Richardson GP (ed) (1988) *System Dynamics Review*. Chaos Special Issue 4:1–2
49. Richmond B (1980) A new look at an old friend. Plexus, Resource Policy Center, Thayer School of Engineering, Dartmouth College, Hanover
50. Saleh M (2002) The characterization of model behavior and its causal foundation. Ph D Thesis, Dept of Information Science, University of Bergen, Bergen
51. Saleh M, Davidsen P (2001) The origins of behavior patterns. *Proceedings of the Int System Dynamics Conference*, Atlanta, July 2001. System Dynamics Society, Albany
52. Saleh M, Davidsen P (2001) The origins of business cycles. *Proceedings of the Int System Dynamics Conference*, Atlanta, July 2001. System Dynamics Society, Albany
53. Saleh M, Oliva R, Davidsen P, Kampmann CE (2006) Eigenvalue Analysis of System Dynamics Models: Another Perspective. *Proceedings of the Int System Dynamics Conference*, Neijmegen, The Netherlands, July 2006. System Dynamics Society, Albany
54. Saleh M, Oliva R, Davidsen P, Kampmann CE (2008) A comprehensive analytical approach for policy analysis of system dynamics models. Mays Business School, Texas A&M University, Working Paper
55. Schweppe F (1973) *Uncertain Dynamical Systems*. Prentice-Hall, Englewood Cliffs
56. Senge PM (1990) *The Fifth Discipline: The Art & Practice of the Learning Organization*. Doubleday Currency, New York
57. Sterman JD (1981) *The Energy Transition and the Economy: A System Dynamics Approach*. Ph D Thesis, Sloan School of Management, MIT, Cambridge
58. Sterman JD (1985) A Behavioral Model of the Economic Long Wave. *J Econ Behav Org* 6(1):17–53
59. Sterman JD (2000) *Business dynamics: Systems thinking and modeling for a complex world*. Irwin McGraw-Hill, Boston
60. Wolstenholme E (2004) Using generic system archetypes to support thinking and modelling. *Syst Dyn Rev* 20(4):341–356
61. Zhushubaliyev ZT, Mosekilde E (2003) *Bifurcation and Chaos in Piecewise-Smooth Dynamical Systems*. World Scientific, Singapore

## System Dynamics, The Basic Elements of

GEORGE P. RICHARDSON

Rockefeller College of Public Affairs and Policy,  
University at Albany, State University of New York,  
Albany, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Feedback Thinking](#)

[Loop Dominance and Nonlinearity](#)

[The Endogenous Point of View](#)

[System Structure](#)

[Suggestions for Further Reading  
on the Core of System Dynamics](#)

[Bibliography](#)

### Glossary

**Endogenous** Generated from within. Contrasting with “exogenous,” meaning generated by forces external to a system or point of view.

**Feedback loop** A closed path of causal influences and information, forming a circular-causal loop of information and action.

**System dynamics** System dynamics is a computer-aided approach to theory-building, policy analysis and strategic decision support emerging from an endogenous point of view.

### Definition of the Subject

System dynamics is a computer-aided approach to theory-building, policy analysis, and strategic decision support emerging from an endogenous point of view [18,20]. It applies to dynamic problems arising in complex social, managerial, economic, or ecological systems – literally any dynamic systems characterized by interdependence, mutual interaction, information feedback, and circular causality.

### Introduction

The field of system dynamics developed initially from the work of Jay W. Forrester. His seminal book *Industrial Dynamics* [7] is still a significant statement of philosophy and methodology in the field. Within ten years of its publication, the span of applications grew from corporate and industrial problems to include the management of research

and development, urban stagnation and decay, commodity cycles, and the dynamics of growth in a finite world. It is now applied in economics, public policy, environmental studies, defense, theory-building in social science, and other areas, as well as its home field, management. The name industrial dynamics no longer does justice to the breadth of the field (for extensive examples, see [20,28], so it has become generalized to system dynamics. The modern name suggests links to other systems methodologies, but the links are weak and misleading. System dynamics emerges out of servomechanisms engineering, not general systems theory or cybernetics [18].

The system dynamics approach involves:

- Defining problems dynamically, in terms of graphs over time.
- Striving for an endogenous, behavioral view of the significant dynamics of a system, a focus inward on the characteristics of a system that themselves generate or exacerbate the perceived problem.
- Thinking of all concepts in the real system as continuous quantities interconnected in loops of information feedback and circular causality.
- Identifying independent stocks or accumulations (levels) in the system and their inflows and outflows (rates).
- Formulating a behavioral model capable of reproducing, by itself, the dynamic problem of concern. The model is usually a computer simulation model expressed in nonlinear equations, but is occasionally left unquantified as a diagram capturing the stock-and-flow/causal feedback structure of the system.
- Deriving understandings and applicable policy insights from the resulting model.
- Implementing changes resulting from model-based understandings and insights.

Mathematically, the basic structure of a formal system dynamics computer simulation model is a system of coupled, nonlinear, first-order differential (or integral) equations,

$$\frac{d}{dt}\mathbf{x}(t) = \mathbf{f}(\mathbf{x}, \mathbf{p}),$$

where  $\mathbf{x}$  is a vector of levels (stocks or state variables),  $\mathbf{p}$  is a set of parameters, and  $\mathbf{f}$  is a nonlinear vector-valued function. Such a system has been variously called a *state-determined system* in the engineering literature, an *absolute system* [3], an *equifinal system* [32], and a *dynamical system* [16].

Simulation of such systems is easily accomplished by partitioning simulated time into discrete intervals of length  $dt$  and stepping the system through time one  $dt$

at a time. Each state variable is computed from its previous value and its net rate of change  $x'(t)$ :  $x(t) = x(t - dt) + dt \cdot x'(t - dt)$ . In the earliest simulation language in the field (DYNAMO) this equation was written with time scripts K (the current moment), J (the previous moment), and JK (the interval between time J and K):  $X_K = X_J + DT \cdot XRATE_{JK}$  (see, e. g., [22]). The computation interval  $dt$  is selected small enough to have no discernible effect on the patterns of dynamic behavior exhibited by the model. In more recent simulation environments, more sophisticated integration schemes are available (although the equation written by the user may look like this simple Euler integration scheme), and time scripts may not be in evidence. Important current simulation environments include STELLA and iThink (isee Systems, <http://www.iseesystems.com/>), Vensim (Ventana Systems, <http://www.vensim.com/>), and Powersim (<http://www.powersim.com/>).

Forrester's original work stressed a continuous approach, but increasingly modern applications of system dynamics contain a mix of discrete difference equations and continuous differential or integral equations. Some practitioners associated with the field of system dynamics work on the mathematics of such structures, including the theory and mechanics of computer simulation, analysis and simplification of dynamic systems, policy optimization, dynamical systems theory, and complex nonlinear dynamics and deterministic chaos.

The main applied work in the field, however, focuses on understanding the dynamics of complex systems for the purpose of policy analysis and design. The conceptual tools and concepts of the field – including feedback thinking, stocks and flows, the concept of feedback loop dominance, and an endogenous point of view – are as important to the field as its simulation methods.

## Feedback Thinking

Conceptually, the feedback concept is at the heart of the system dynamics approach. Diagrams of loops of information feedback and circular causality are tools for conceptualizing the structure of a complex system and for communicating model-based insights. Intuitively, a feedback loop exists when information resulting from some action travels through a system and eventually returns in some form to its point of origin, potentially influencing future action. If the tendency in the loop is to reinforce the initial action, the loop is called a *positive* or *reinforcing* feedback loop; if the tendency is to oppose the initial action, the loop is called a *negative*, *counteracting*, or *balancing* feedback loop. The sign of the loop is called its *po-*

larity. Balancing loops can be variously characterized as goal-seeking, equilibrating, or stabilizing processes. They can sometimes generate oscillations, as when a pendulum seeking its equilibrium goal gathers momentum and overshoots it. Reinforcing loops are sources of growth or accelerating collapse; they are disequilibrating and destabilizing. Combined, balancing and reinforcing circular causal feedback loops can generate all manner of dynamic patterns.

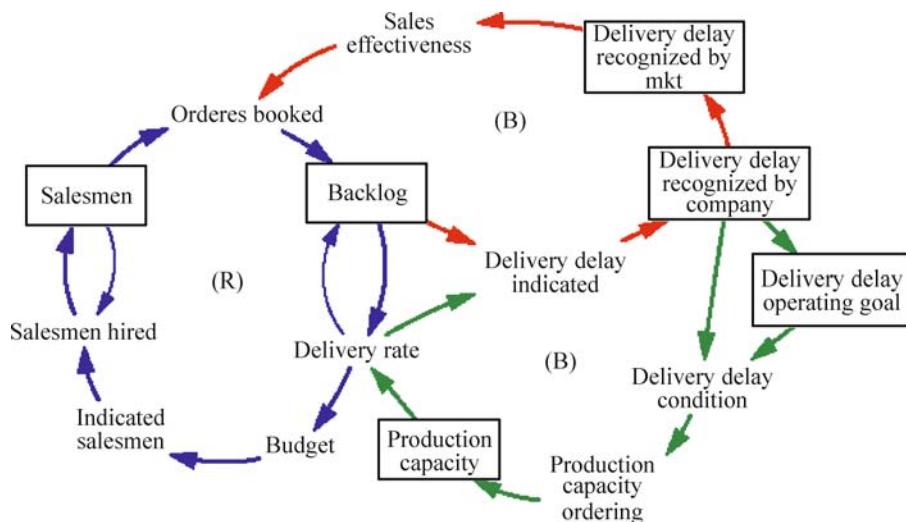
Feedback loops are ubiquitous in human and natural systems and, under various names and representations, have been widely recognized in popular and scholarly literature. Feedback thought has been present implicitly or explicitly for hundreds of years in the social sciences and literally thousands of years in recorded history [9]. We have the vicious circle originating in classical logic and morphing into common usage, the bandwagon effect, the invisible hand of Adam Smith, Malthus's correct observation of population growth as a self-reinforcing process, Keynes's consumption multiplier, the investment accelerator of Hicks and Samuelson, compound interest or inflation, the biological concepts of proprioception and homeostasis, Festinger's cognitive dissonance, Myrdal's principle of cumulative causation, Venn's idea of a suicidal prophecy, Merton's related notion of a self-fulfilling prophecy, and so on. Each of these ideas can be concisely and insightfully represented as one or more loops of causal influences with positive or negative polarities. Great social scientists and feedback thinkers; great social theories are

feedback thoughts. (For a full exposition of the evolution of the feedback concept see [19].)

### Loop Dominance and Nonlinearity

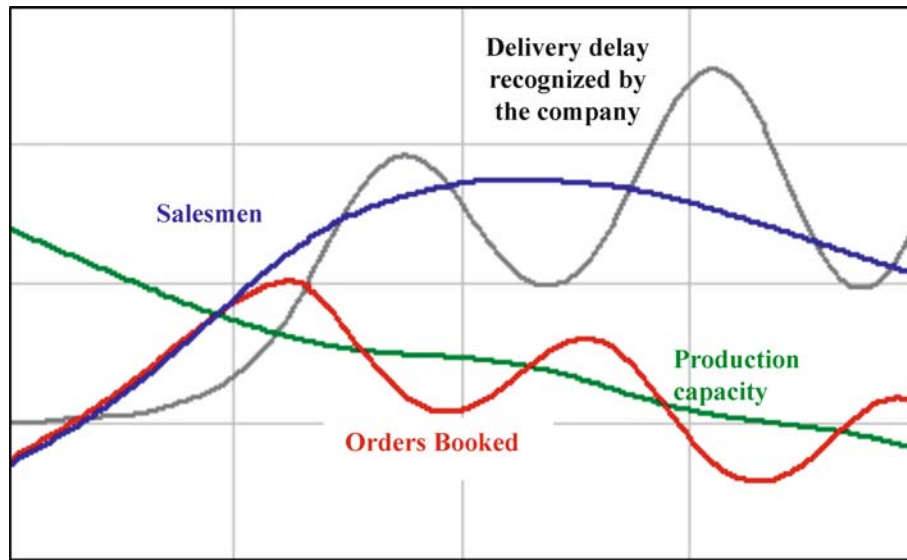
The loop concept underlying feedback and circular causality by itself is not enough, however. The explanatory power and insightfulness of feedback understandings also rest on the notions of active structure and loop dominance. Complex systems change over time. A crucial requirement for a powerful view of a dynamic system is the ability of a mental or formal model to change the strengths of influences as conditions change, that is to say, the ability to shift *active* or *dominant structure*.

In a system of equations, this ability to shift loop dominance comes about endogenously from nonlinearities in the system. For example, the S-shaped dynamic behavior of the classic logistic growth model ( $dP/dt = aP - bP^2$ ) or similar structures like the Gompertz curve ( $dP/dt = aP - bP \ln(P)$ ) can be seen as the consequence of a shift in loop dominance from a positive, self-reinforcing feedback loop ( $aP$ ) producing exponential-like growth, to a negative feedback loop ( $-bP^2$  or  $-bP \ln(P)$ ) that brings the system to its eventual goal. The shift in loop dominance in these models comes about from the nonlinearity in the second term, which grows faster than the first term and eventually overtakes it. Only nonlinear models can endogenously alter their active or dominant structure and shift loop dominance.



System Dynamics, The Basic Elements of, Figure 1

Core structure of Forrester's market growth model [8], showing a blue reinforcing loop underlying the growth (or reinforcing decline) of Salesmen, Orders, and Revenue, a red balancing loop containing various delayed recognitions of the company's delivery delay, and a green balancing loop responsible for capacity ordering if the delivery delay drops too far below its operating goal



**System Dynamics, The Basic Elements of, Figure 2**

The dynamic behavior of the model shown in Fig. 1, illustrating an early growth phase, which turns into an oscillatory phase as the feedback loop dominance shifts to the *red balancing delivery delay loop*, and results in a long term corporate decline as the *green capacity ordering loop* responds to a sliding operating goal for the acceptable delivery delay

Real systems are perceived to change their active or dominant structure over time, often because of the build-up of internal forces. Thus from a feedback perspective, the ability of nonlinearities to generate shifts in loop dominance is the fundamental reason for advocating nonlinear models of social system behavior.

Figures 1 and 2, abstracted from an early, classic paper [8] illustrate these ideas. In Fig. 1 salesmen (in the blue reinforcing loop) book orders for the company; if enough revenue is generated, there is enough budget to hire more salesmen and corporate growth ensues. Whether salesmen (in this simplified picture) book enough orders depends on the company's delivery delay for the product, as perceived by the market (red balancing loop). The company builds production capacity according to its perceived need, as indicated by its perceived delivery delay and its target for that (green balancing loop).

Figure 2 shows the dynamics this feedback structure endogenously generates. In the early phase, salesmen grow as orders and revenue grow; the system's exponential growth behavior in that phase is generated by the reinforcing salesmen loop. But then the feedback loop dominance soon shifts to the balancing delivery delay loop, which constrains sales effectiveness and brings a halt to growth. The system moves into an oscillatory phase generated by the various monitoring and perception delays around the now dominant red balancing loop. Salesmen eventual peak and decline, as the green production capacity ordering loop

fails to keep production capacity sufficient to hold the delivery delays in check.

Thus the dynamic behavior of this system is a consequence of its feedback structure and the nonlinearities that shift loop dominance endogenously over time. The particular decline scenario shown in Fig. 2 illustrates one of the deep insights of the model: the adaptive goal structure, in which the delivery delay operating goal moves slowly to accommodate changes in the company's delivery delay, weakens the green balancing loop trying to bring on capacity. The company never perceives its delivery delay is sufficiently higher than its (sliding) target, so it fails to order sufficient capacity to sustain growth. A fixed goal for the acceptable delivery delay sends a stronger signal, which can turn this corporate decline into oscillating growth [8].

Thus, nonlinearity is crucial to the system dynamics approach. However, it is crucial not merely because of its mathematical properties but because it enables the formalization of a profoundly powerful perspective on theory and policy – the *endogenous point of view*.

### The Endogenous Point of View

The concept of endogenous change is fundamental to the system dynamics approach. It has both philosophical and engineering origins. A deep and lasting insight of the earliest attempts at servomechanisms control is the realization that *the attempt to control a system generates dynamics of*



*its own*, complicating the dynamics trying to be controlled. A governor mechanism imposed to control the speed of a steam engine can generate oscillatory “hunting behavior,” as the control system overshoots and undershoots the set point. As it becomes part of the system, the governing mechanism thus generates dynamics of its own.

The insight transfers readily, but with added significance, from engineering systems to people systems: Attempts to control complex human systems – coercing, guiding, managing, governing – generate dynamics of their own. Moreover, some of these endogenously generated dynamics are created by the control mechanisms themselves (like the governor of a steam engine) and some are created by human creative responses to the management efforts (e.g., principal-agent interactions). These natural and human forces, creating counteracting and compensating pressures in response to system control efforts, emerge as complicated circular-causal feedback structures. The often complex, difficult-to-understand dynamics of such management systems are to a great degree a consequence of their internal structures.

To capture and analyze such management complexities, one must look inward to see the ways a complex system naturally responds to system pressures. The endogenous point of view is thus central to the system dynamics approach. It dictates aspects of model formulation: exogenous disturbances are seen at most as *triggers* of system behavior (like displacing a pendulum); the *causes* are contained within the structure of the system itself (like the interaction of a pendulum’s position and momentum that produces oscillations). Corrective responses are also not modeled as functions of time, but are dependent on conditions within the system. Time by itself is not seen as a cause in the endogenous point of view.

Theory building and policy analysis are significantly affected by this endogenous perspective. Taking an endogenous view exposes the natural *compensating* tendencies in social systems that conspire to defeat many policy initiatives. Feedback and circular causality are delayed, devious, and deceptive. For understanding, system dynamics practitioners strive for an *endogenous point of view*. The effort is to uncover the sources of system behavior that exist within the structure of the system itself.

## System Structure

These ideas are captured almost explicitly in Forrester’s [9] organizing framework for system structure:

- Rates
- Goal
- Observed condition
- Discrepancy
- Desired action.

The *closed boundary* signals the endogenous point of view. The word *closed* here does not refer to open and closed systems in the general system sense, but rather refers to the effort to view a system as *causally* closed. The modeler’s goal is to assemble a formal structure that can, *by itself*, without exogenous explanations, reproduce the essential characteristics of a dynamic problem.

The causally closed system boundary at the head of this organizing framework identifies the endogenous point of view as the feedback view pressed to an extreme. Feedback thinking can be seen as a *consequence* of the effort to capture dynamics within a closed causal boundary. Without causal loops, all variables must trace the sources of their variation ultimately outside a system. Assuming instead that the causes of all significant behavior in the system are contained within some closed causal boundary forces causal influences to feed back upon themselves, forming causal loops. Feedback loops enable the endogenous point of view and give it structure.

## Levels and Rates

Stocks (accumulations, or “levels” in early system dynamics literature) and the flows (“rates”) that affect them are essential components of system structure. A map of causal influences and feedback loops is not enough to determine the dynamic behavior of a system. A constant inflow yields a linearly rising stock; a linearly rising inflow yields a stock rising along a parabolic path; a stock with inflow proportional to itself grows exponentially; two stocks in a balancing loop have a tendency to generate oscillations; and so on. For example, the boxes in Fig. 1 represent accumulations in the company and its market; the three stocks in the red balancing loop (the order backlog and the two perceptions of the company’s delivery delay) give that loop its tendency to generate oscillations which propagate through out the system. Accumulations are the memory of a dynamic system and contribute to its disequilibrium and dynamic behavior.

Forrester [7] placed the operating policies of a system among its rates, the inflows and outflows governing change in the system. Many of these rates of change assume the classic structure of a negative feedback loop striving to take action to reduce the discrepancy between the observed condition of the system and a goal. The simplest

- Closed boundary
- Feedback loops
- Levels

such rate structure results in an equation of the form

$$\text{RATE} = \frac{\text{GOAL} - \text{LEVEL}}{\text{ADJUSTMENT TIME}},$$

where ADJUSTMENT TIME is the time over which the level adjusts to reach the goal. This simple formulation reflects Forrester's more general statement about rates in his hierarchy of system structure (above) which can be richly thought of as

$$\text{RATE} = f(\text{DESIRED ACTION})$$

$$\begin{aligned} \text{DESIRED ACTION} \\ = g(\text{DESIRED CONDITION}, \\ \text{OBSERVED CONDITION}) \end{aligned}$$

$$\text{OBSERVED CONDITION} = h(\text{LEVELS}),$$

for some functions  $f$ ,  $g$ , and  $h$  representing particular system characteristics.

Operating policies in a management system can influence the *flows* of information, material, and resources, which are the only means of changing the accumulations in the system. While flows can be changed quickly, as a matter of relatively quick decision making, stocks change slowly – they rise when inflows are great than outflows, and decline when inflows are less than outflows.

The simple “tub dynamics” of stocks are clear even to children, yet can be befuddling in complex systems. The accumulation of green house gases in the atmosphere, for example, affects the *flow* of heat energy radiated from the earth. To turn around global warming, the *accumulation* of green house gases must drop far enough to raise radiant energy above the inflow of solar energy, a simple stock-and-flow insight. But to cause the accumulation of green house gases to drop, their generation must fall below their natural absorption rate (another simple stock-and-flow observation). So turning around global warming is a process involving a chain of at least two significant accumulations, and people have trouble thinking it through reliably. The accumulations can only be changed by managing their associated flows. They will change only slowly even if we manage the technical and political pitfalls involved in lowering green house gas production (see [29]).

The significance of stocks in complex systems is vivid in a resource-based view of strategy and policy. Resources that enable a corporation or government to function or flourish are stocks, usually accumulated over long periods of time with significant investment of time, energy, and money. Reputations are also stocks, built over similarly

long periods of time. While inadequate by themselves to give a full picture of the dynamics of a complex system, stocks and flows are vital components of system structure, without which fundamental understandings of dynamics are impossible [33].

### Behavior is a Consequence of System Structure

The importance of stocks and flows appears most clearly when one takes a *continuous* view of structure and dynamics. Although a discrete view, focusing on separate events and decisions, is entirely compatible with an endogenous feedback perspective, the system dynamics approach emphasizes a continuous view [7]. The continuous view strives to look beyond events to see the dynamic patterns underlying them: model not the appearance of a discrete new housing unit in a city, but focus instead on the rise and fall of aggregate numbers of housing units. Moreover, the continuous view focuses not on discrete decisions but on the *policy structure* underlying decisions: not why this particular apartment building was constructed but what persistent pressures exist in the urban system that produce decisions that change housing availability in the city. Events and decisions are seen as surface phenomena that ride on an underlying tide of system structure and behavior. It is that underlying tide of policy structure and continuous behavior that is the system dynamicist's focus.

There is thus a *distancing* inherent in the system dynamics approach – not so close as to be confused by discrete decisions and myriad operational details, but not so far away as to miss the critical elements of policy structure and behavior. Events are deliberately blurred into dynamic behavior. Decisions are deliberately blurred into perceived policy structures. Insights into the connections between system structure and dynamic behavior, which are the goal of the system dynamics approach, come from this particular distance of perspective.

### Suggestions for Further Reading on the Core of System Dynamics

The *System Dynamics Review*, the journal of the System Dynamics Society, published by Wiley, is the best source of current activity in the field, including methodological advances and applications.

The core of a vibrant field is difficult to discern in the flow of current work. However, the works that the field itself singles out as exemplary can give some reliable hints about what is considered vital to the core. In this sense two edited volumes are noteworthy: An early, interesting collection of applications is Roberts [24]; Richardson [21] is a more recent two-volume edited collection in

the same spirit, containing prize-winning work in philosophical background, dynamic decision making, applications in the private and public sectors, and techniques for modeling with management.

In addition, the following works, selected from among winners of the System Dynamics Society's *Jay Wright Forrester Award* (see [www.systemdynamics.org/Society\\_Awards.htm](http://www.systemdynamics.org/Society_Awards.htm)), can be considered insightful although implicit exemplars of the core of system dynamics. (Publications are listed beginning with the most recent; see the bibliography for full citations):

- Thomas S. Fiddaman, "Exploring policy options with a behavioral climate-economy model"
- Kim D. Warren, *Competitive Strategy Dynamics*
- Eric F. Wolstenholme, "Towards the Definition and Use of a Core Set of Archetypal Structures in System Dynamics"
- Nelson P. Repenning, "Understanding Fire Fighting in New Product Development"
- John D. Sterman, *Business Dynamics, Systems Thinking and Modeling for a Complex World*
- Peter Milling, "Modeling innovation processes for decision support and management simulation."
- Erling Moxnes, "Not Only the Tragedy of the Commons: Misperceptions of Bioeconomics."
- Jac A. M. Vennix, *Group Model Building: Facilitating Team Learning Using System Dynamics*
- Jack B. Homer, "A System Dynamics Model of National Cocaine Prevalence."
- Andrew Ford, "Estimating the Impact of Efficiency Standards on Uncertainty of the Northwest Electric System."
- Khalid Saeed, *Towards Sustainable Development: Essays on System Analysis of National Policy*
- Tarek Abdul-Hamid and Stuart Madnick, *Software Project Dynamics: An Integrated Approach*
- George P. Richardson, *Feedback Thought in Social Science and Systems Theory*
- Peter M. Senge, *The Fifth Discipline*
- John D. W. Morecroft, "Rationality in the Analysis of Behavioral Simulation Models."
- John D. Sterman, "Modeling Managerial Behavior: Misperceptions of Feedback in a Dynamic Decision Making Experiment."

For texts on the system dynamics approach, see Alfeld and Graham [2], Richardson and Pugh [22], Wolstenholme [34], Ford [6], Maani and Cavana [11], and the most comprehensive text to date, Sterman [28].

## Bibliography

1. Abdul-Hamid T, Madnick S (1991) *Software project dynamics: an integrated approach*. Prentice Hall, Englewood Cliffs
2. Alfeld LE, Graham AK (1976) *Introduction to urban dynamics*. Pegasus Communications, Waltham
3. Ashby H (1956) *An Introduction to cybernetics*. Chapman & Hall, London
4. Fiddaman TS (2002) Exploring policy options with a behavioral climate-economy model. *Syst Dyn Rev* 18(2):243–267
5. Ford A (1990) Estimating the impact of efficiency standards on uncertainty of the northwest electric system. *Oper Res* 38(4):580–597
6. Ford A (1999) *Modeling the environment: an introduction to system dynamics of environmental systems*. Island Press, Washington
7. Forrester JW (1961) *Industrial dynamics*. MIT Press, Cambridge. Reprinted by Pegasus Communications, Waltham
8. Forrester JW (1968) Market growth as influenced by capital investment. *Ind Manag Rev* (MIT, now *Sloan Manag Rev*) 9(2):83–105. Reprinted widely, e. g., Richardson [21]
9. Forrester JW (1969) *Urban dynamics*. MIT Press, Cambridge. Reprinted by Pegasus Communications, Waltham
10. Homer JB (1992) A system dynamics model of national cocaine prevalence. *Syst Dyn Rev* 9(1):49–78
11. Maani KE, Cavana RY (2000) *Systems thinking and modelling: understanding change and complexity*. Pearson Education, New Zealand
12. Milling P (1996) Modeling innovation processes for decision support and management simulation. *Syst Dyn Rev* 12(3):211–234
13. John DW, Morecroft JDW (1985) Rationality in the analysis of behavioral simulation models. *Manag Sci* 31(7):900–916
14. Morecroft JDW, Sterman JD (eds) (1994) *Modeling for learning organizations*. System dynamics series. Pegasus Communications, Waltham
15. Moxnes E (1998) Not only the tragedy of the commons: misperceptions of bioeconomics. *Manag Sci* 44(9):1234–1248
16. Nicholas G, Prigogine I (1977) *Self-organization in nonequilibrium systems: from dissipative structures to order through fluctuations*. Wiley, New York
17. Repenning NR (2001) Understanding fire fighting in new product development. *J Prod Innov Manag* 18(5):285–300
18. Richardson GP (1991) System dynamics: simulation for policy analysis from a feedback perspective. In: Fishwick PA, Luker PA (eds) *Qualitative simulation modeling and analysis*. Springer, New York
19. Richardson GP (1991) *Feedback thought in social science and systems theory*. University of Pennsylvania Press, Philadelphia. Reprinted by Pegasus Communications, 1999
20. Richardson GP (1996) System dynamics. In: Gass S, Harris C (eds) *The encyclopedia of operations research and management science*. Kluwer, New York
21. Richardson GP (ed) (1996) *Modelling for management: simulation in support of systems thinking*. International library of management. Dartmouth, Aldershot
22. Richardson GP, Pugh AL III. (1981) *Introduction to system dynamics modeling with DYNAMO*. MIT Press, Cambridge. Reprinted by Pegasus Communications, Waltham
23. Richmond B (1993) Systems thinking: critical thinking skills for the 1990s and beyond. *Syst Dyn Rev* 9(2):113–133

24. Roberts EB (ed) (1978) Managerial applications of system dynamics. MIT Press, Cambridge. Reprinted by Pegasus Communications, Waltham
25. Saeed K (1991) Towards sustainable development: essays on system analysis of national policy. Progressive Publishers, Lahore
26. Senge PM (1990) The fifth discipline: the art and practice of the learning organization. Doubleday/Currency, New York
27. Sterman JD (1988) Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment. *Manag Sci* 35(3):321–339
28. Sterman JD (2001) Business dynamics: systems thinking and modeling for a complex world. Irwin McGraw-Hill, Boston
29. Sterman JD, Sweeney LB (2002) Cloudy skies: assessing public understanding of global warming. *Syst Dyn Rev* 18(2):207–240
30. System Dynamics Review. 1985–present. Wiley, Chichester
31. Vennix JAM (1996) Group model building: facilitating team learning using system dynamics. Wiley, Chichester
32. von Bertalanffy L (1968) General systems theory: foundations, development, applications. George Braziller, New York
33. Warren KD (2002) Competitive strategy dynamics. Wiley, United Kingdom
34. Wolstenholme EF (1990) System enquiry: a system dynamics approach. Wiley, Chichester
35. Wolstenholme EF (2003) Towards the definition and use of a core set of archetypal structures in system dynamics. *Syst Dyn Rev* 19(1):7–26

## System Dynamics in the Evolution of the Systems Approach

MARKUS SCHWANINGER

Institute of Management, University of St. Gallen,  
St. Gallen, Switzerland

### Article Outline

Glossary

Definition of the Subject

Introduction

Emergence of the Systems Approach

Common Grounds and Differences

The Variety of Systems Methodologies

System Dynamics –

Its Features, Strengths and Limitations

Actual and Potential Relationships

Outlook

Appendix

Bibliography

### Glossary

**Cybernetics** The science of communication and control in complex, dynamical systems. The core objects of study are information, communication, feedback and

adaptation. In the newer versions of cybernetics, the emphasis is on observation, self-organization, self-reference and learning.

**Dynamical system** The dynamical system concept is a mathematical formalization of time-dependent processes. Examples include the mathematical models that describe the swinging of a clock pendulum, the flow of water in a river, and the evolution of a population of fish in a lake.

**Law of requisite variety** Ashby's law of requisite variety says: "Only variety can destroy variety". It implies that the varieties of two interacting systems must be in balance, if stability is to be achieved.

**Organizational cybernetics** The science which applies cybernetic principles to organization. Synonyms are *Management Cybernetics* and *Managerial Cybernetics*.

**System** There are many definitions of *system*. Two examples: A portion of the world sufficiently well defined to be the subject of study; something characterized by a structure, for example, a social system (Anatol Rapoport). A system is a family of relationships between its members acting as a whole (International Society for the Systems Sciences).

**System dynamics** A methodology and discipline for the modeling, simulation and control of dynamic systems. The main emphasis falls on the role of structure and its relationship with the dynamic behavior of systems, which are modeled as networks of informationally closed feedback loops between stock and flow variables.

**Systems approach** A perspective of inquiry, education and management, which is based on system theory and cybernetics.

**System theory** A formal science of the structure, behavior, and development of systems. In fact there are different system theories. General system theory is a transdisciplinary framework for the description and analysis of any kind of system. System theories have been developed in many domains, e.g., mathematics, computer science, engineering, sociology, psychotherapy, biology and ecology.

**Variety** A technical term for *complexity* which denotes the number of (potential) states of a system.

### Definition of the Subject

The purpose of this chapter is to give an overview of the role of system dynamics (SD) in the context of the evolution of the systems movement. This is necessary because SD is often erroneously taken as the systems approach as such, not as part of it. It is also requisite to show that the



processes of the evolution of both SD in particular and the systems movement as a whole are intimately linked and intertwined. Finally, in view of the purpose of the chapter the actual and potential relationships between system dynamics and the other strands of the systems movement are evaluated. This way, complementarities and synergies are identified.

## Introduction

The purpose of this contribution is to give an overview of the role of system dynamics in the context of the evolution of the systems movement. “Systems movement” – often referred to briefly as “systemics” – is a broad term, which takes into account the fact that there is no single system approach, but a range of different ones. The common denominator of the different system approaches in our day is that they share a worldview focused on complex dynamic systems, and an interest in describing, explaining and designing or at least influencing them. Therefore, most of the system approaches offer not only a theory but also a way of thinking (“systems thinking” or “systemic thinking”) and a methodology for dealing with systemic issues or problems.

System dynamics (SD) is a discipline and a methodology for the modeling, simulation and control of complex, dynamic systems. SD was developed by MIT professor Jay W. Forrester (e.g. [20,21]) and has been propagated by his students and associates. SD has grown to a school of numerous academics and practitioners all over the world. The particular approach of SD lies in representing the issues or systems-in-focus as meshes of closed feedback loops made up of stocks and flows, in continuous time and subject to delays.

The development of the system dynamics methodology and the worldwide community that applies SD to modeling and simulation in radically different contexts suggest that it is a “systems approach” on its own. Nevertheless, taking “system dynamics” as the (one and only) synonym for “systemic thinking” would be going too far, given the other approaches to systemic thinking as well as a variety of system theories and methodologies, many of which are complementary to SD. In any case, however, the SD community has become the strongest “school” of the Systems approach, if one takes the numbers of members in organizations representing the different schools as a measure (by 2006, the System Dynamics Society had more than 1000 members).

The rationale and structure of this contribution is as follows. Starting with the emergence of the systems approach, the multiple roots and theoretical streams of sys-

temics are outlined. Next, the common grounds and differences among different strands of the systems approach are highlighted, and the various systems methodologies are explored. Then the distinctive features of SD are analyzed. Finally comes a reflection on the relationships of SD with the rest of the systems movement as well as with potential complementarities and synergies.

In Table 1, a time-line overview of some milestones in the evolution of the systems approach in general and System Dynamics in particular is given. Elaborating on each of the sources quoted therein would reach beyond the purpose of this chapter. However, to convey a synoptic view, a diagram showing the different systems approaches and their interrelationships is provided in the Appendix “[Systems Approaches – An Overview](#)”.

## Emergence of the Systems Approach

The systems movement has many roots and facets, with some of its concepts going back as far as ancient Greece. What we name as “the systems approach” today materialized in the first half of the twentieth century. At least two important components should be mentioned: those proposed by von Bertalanffy and by Wiener.

Ludwig von Bertalanffy, an American biologist of Austrian origin, developed the idea that organized wholes of any kind should be describable and, to a certain extent, explainable, by means of the same categories, and ultimately by one and the same formal apparatus. His *general systems theory* triggered a whole movement which has tried to identify invariant structures and mechanisms across different kinds of organized wholes (for example, hierarchy, teleology, purposefulness, differentiation, morphogenesis, stability, ultrastability, emergence, and evolution).

In 1948 Norbert Wiener, an American mathematician at the Massachusetts Institute of Technology, published his seminal book on *Cybernetics*, building upon interdisciplinary work carried out in cooperation with Bigelow, an IBM engineer, and Rosenblueth, a physiologist. Wiener’s opus became the transdisciplinary foundation for a new science of capturing as well as designing control and communication mechanisms in all kinds of dynamic systems [81]. Cyberneticists have been interested in concepts such as information, communication, complexity, autonomy, interdependence, cooperation and conflict, self-production (“autopoiesis”), self-organization, (self-) control, self-reference and (self-) transformation of complex dynamic systems.

Along the genetic line of the tradition which led to the evolution of General Systems Theory (von Berta-



lanffy, Boulding, Gerard, Miller, Rapoport) and Cybernetics (Wiener, McCulloch, Ashby, Powers, Pask, Beer), a number of roots can be identified, in particular:

- Mathematics (for example, Newton, Poincaré, Lyapunov, Lotka, Volterra, Rashevsky)
- Logic (for example, Epimenides, Leibniz, Boole, Russell and Whitehead, Goedel, Spencer-Brown)
- Biology, including general physiology and neurophysiology (for example, Hippocrates, Cannon, Rosenblueth, McCulloch, Rosen)
- Engineering and computer science, including the respective physical and mathematical foundations (for example, Heron, Kepler, Watt, Euler, Fourier, Maxwell, Hertz, Turing, Shannon and Weaver, von Neumann, Walsh)
- Social and human sciences, including economics (for example, Hume, Adam Smith, Adam Ferguson, John Stuart Mill, Dewey, Bateson, Merton, Simon, Piaget).

In this last-mentioned strand of the systems movement, one focus of inquiry is on the role of feedback in communication and control in (and between) organizations and society, as well as in technical systems. The other focus of interest is on the multidimensional nature and the multi-level structures of complex systems. Specific theory building, methodological developments and pertinent applications have occurred at the following levels:

- Individual and family levels (for example, systemic psychotherapy, family therapy, holistic medicine, cognitive therapy, reality therapy)
- Organizational and societal levels (for example, managerial cybernetics, organizational cybernetics, socio-cybernetics, social systems design, social ecology, learning organizations)
- The level of complex (socio-)technical systems (systems engineering)

The notion of “socio-technical systems” has become widely used in the context of the design of organized wholes involving interactions of people and technology (for instance, Linstone’s multi-perspectives-framework, known by way of the mnemonic TOP (Technical, Organizational, Personal/individual)).

As can be noted from these preliminaries, different kinds of system theory and methodology have evolved over time. One of these is a theory of dynamic systems by Jay W. Forrester, which serves as a basis for the methodology of system dynamics. Two eminent titles are [20] and [21]. In SD, the main emphasis falls on the role of structure and its relationship with the dynamic behavior of systems, modeled as networks of informationally closed

feedback loops between stock and flow variables. Several other mathematical systems theories have been elaborated, for example, mathematical general systems theory (Klir, Pestel, Mesarovic and Takahara), as well as a whole stream of theoretical developments which can be subsumed under the terms “dynamic systems theory” or “theories of non-linear dynamics” (for example, catastrophe theory, chaos theory and complexity theory). Under the latter, branches such as the theory of fractals (Mandelbrot), geometry of behavior (Abraham), self-organized criticality (Bak), and network theory (Barabasi, Watts) are subsumed. In this context, the term “sciences of complexity” is used.

In addition, a number of mathematical theories, which can be called “system theories,” have emerged in different application contexts, examples of which are discernible in the following fields:

- Engineering, namely information and communication theory (Shannon and Weaver), technology and computer-aided systems theory (for example, control theory, automata, cellular automata, agent-based modeling, artificial intelligence, cybernetic machines, neural nets)
- Operations research (for example, modeling theory and simulation methodologies, Markov chains, genetic algorithms, fuzzy control, orthogonal sets, rough sets)
- Social sciences, economics in particular (for example, game theory, decision theory)
- Biology (for example, Sabelli’s Bios theory of creation)
- Ecology (for example, E. and H. Odum’s systems ecology).

Most of these theories are transdisciplinary in nature, i. e., they can be applied across disciplines. The Bios theory, for example is applicable to clinical, social, ecological and personal settings [54]. Examples of essentially non-mathematical system theories can be found in many different areas of study, e. g.:

- Economics, namely its institutional/evolutionist strand (Veblen, Myrdal, Boulding, Dopfer)
- Sociology (for example, Parsons’ and Luhmann’s social system theories, Hall’s cultural systems theory)
- Political sciences (for example, Easton, Deutsch, Wallerstein)
- Anthropology (for example, Levi Strauss’s structuralist-functionalist anthropology, Margaret Mead)
- Semiotics (for example, general semantics (Korzybski, Hayakawa, Rapoport), cybersemiotics (Brier))
- Psychology and psychotherapy (for example, systemic intervention (Bateson, Watzlawick, F. Simon), and fractal affect logic (Ciompi))

- Ethics and epistemology (for example, Vickers, Churchman, von Foerster, van Gigch)

Several system-theoretic contributions have merged the quantitative and the qualitative in new ways. This is the case for example in Rapoport's works in game theory as well as general systems theory, Pask's conversation theory, von Foerster's cybernetics of cybernetics (second-order cybernetics), and Stafford Beer's opus in managerial cybernetics. In all four cases, mathematical expression is virtuously connected to ethical, philosophical, and epistemological reflection. Further examples are Prigogine's theory of dissipative structures, Mandelbrot's theory of fractals, complex adaptive systems (Holland et al.), Kauffman's complexity theory, and Haken's synergetics, all of which combine mathematical analysis and a strong component of qualitative interpretation.

A large number of systems methodologies, with the pertinent threads of systems practice, have emanated from these theoretical developments. Many of them are expounded in detail in specialized encyclopedias (e.g., [27] and, under a specific theme, named *Systems Science and Cybernetics*, of the Encyclopedia of Life Support Systems [18]). In this chapter, only some of these will be addressed explicitly, in order to shed light on the role of SD as part of the systems movement.

### Common Grounds and Differences

Even though the spectrum of system theories and methodologies outlined in the preceding section may seem multifarious, all of them have a strong common denominator: They build on the idea of systems as organized wholes. An objectivist working definition of a system is that of a whole, the organization of which is made up by interrelationships. A subjectivist definition is that of a set of interdependent variables in the mind of an observer, or, a mental construct of a whole, an aspect that has been emphasized by the position of constructivism. *Constructivism* is a synonym for *second-order cybernetics*. While first-order cybernetics concentrates on regulation, information and feedback, second-order cybernetics focuses on observation, self-organization and self-reference. Heinz von Foerster established the distinction between 'observed systems' for the former and 'observing systems' for the latter [74].

From the standpoint of operational philosophy, a system is, as Rapoport says, "a part of the world, which is sufficiently well defined to be the object of an inquiry or also something, which is characterized by a structure, for example, a production system" [50].

In recent systems theory, the aspect of relationships has been emphasized as the main building block of a sys-

tem, as one can see from a definition published by the International Society for the Systems Sciences (ISSS): "A system is a family of relationships between its members acting as a whole" [63]. Also, purpose and interaction have played an important part in reflections on systems: Systems are conceived, in the words of Forrester [21], as "wholes of elements, which cooperate towards a common goal." Purposeful behavior is driven by internal goals, while purposive behavior rests on a function assigned from the outside. Finally, the aspects of open and closed functioning have been emphasized. Open systems are characterized by the import and export of matter, energy and information. A variant of particular relevance in the case of social systems is the operationally closed system, that is, a system which is self-referential in the sense that its self-production (autopoiesis) is a function of production rules and processes by which order and identity are maintained, and which cannot be modified directly from outside. As we shall see, this concept of operational closure is very much in line with the concept of circularity used in SD.

At this point, it is worth elaborating on the specific differences between two major threads of the systems movement, which are of special interest because they are grounded in "feedback thought" [52]: The cybernetic thread, from which organizational cybernetics has emanated, and the servomechanic thread in which SD is grounded. As Richardson's detailed study shows, the strongest influence on cybernetics came from biologists and physiologists, while the thinking of economists and engineers essentially shaped the servomechanic thread. Consequently, the concepts of the former are more focused on the adaptation and control of complex systems for the purpose of maintaining stability under exogenous disturbances. Servomechanics, on the other hand, and SD in particular, take an endogenous view, being mainly interested in understanding circular causality as the principal source of a system's behavior. Cybernetics is more connected with communication theory, the general concern of which can be summarized as how to deal with randomly varying input. SD, on the other hand, shows a stronger link with engineering control theory, which is primarily concerned with behavior generated by the control system itself, and by the role of nonlinearities. Managerial cybernetics and SD both share the concern of contributing to management science, but with different emphases and with instruments that are different but in principle complementary. Finally, the mathematical foundations are generally more evident in the basic literature on SD than in the writings on organizational cybernetics, in which the formal apparatus underlying model formulation is confined to a small number of publications (e.g., [7,10]),

which are less known than the qualitative treatises. The terms *management cybernetics* and *managerial cybernetics* are used as synonyms for *organizational cybernetics*.

### The Variety of Systems Methodologies

The methodologies that have evolved as part of the systems movement cannot be expounded in detail here. The two epistemological strands in which they are grounded, however, can be identified – the positivist tradition and the interpretivist tradition.

*Positivist tradition* denotes those methodological approaches that focus on the generation of “positive knowledge,” that is, a knowledge based on “positively” ascertained facts. *Interpretivist tradition* denotes those methodological approaches that emphasize the importance of subjective interpretations of phenomena. This stream goes back to Greek art and science of the interpretation and understanding of texts.

Some systems methodologies have been rooted in the positivist tradition, and others in the interpretivist tradition. The differences between the two can be described along the following set of polarities:

- An objectivist versus a subjectivist position
- A conceptual–instrumental versus a communicational/cultural/political rationality
- An inclination to quantitative versus qualitative modeling
- A structuralist versus a discursive orientation.

A positivistic methodological position tends toward the objectivistic, conceptual–instrumental, quantitative and structuralist–functionalist in its approach. An interpretive position, on the other hand, tends to emphasize the subjectivist, communicational, cultural, political, ethical and esthetic—that is, the qualitative and discursive aspects. It would be too simplistic to classify a specific methodology in itself as being “positivistic” or “interpretative”. Despite the traditions they have grown out of, several methodologies have evolved and been reinterpreted or opened to new aspects (see below).

In the following, a sample of systems methodologies will be characterized and positioned in relation to these two traditions, beginning with those in the positivistic strand:

- *“Hard” OR methods.* Operations research (OR) uses a wide variety of mathematical and statistical methods and techniques—for example of optimization, queuing, dynamic programming, graph theory, time series analysis—to provide solutions for organizational and man-

agerial problems, mainly in the operational domains of production and logistics, and in finance.

- *Living systems theory.* In his LST, James Grier Miller [44] identifies a set of 20 necessary components that can be discerned in living systems of any kind. These structural features are specified on the basis of a huge empirical study and proposed as the “critical subsystems” that “make up a living system.” LST has been used as a device for diagnosis and design in the domains of engineering and the social sciences.
- *Viable system model.* To date, Stafford Beer’s VSM is probably the most important product of organizational cybernetics. It specifies a set of management functions and their interrelationships as the sufficient conditions for the viability of any human or social system (see [10]). These are applicable in a recursive mode, for example, to the different levels of an organization. The VSM has been widely applied in the diagnostic mode, but also to support the design of all kinds of social systems. Specific methodologies for these purposes have been developed, for instance for use in consultancy. The term viable system diagnosis (VSD) is also used.

The methodologies and models addressed up to this point have by and large been created in the positivistic tradition of science. Other strands in this tradition do exist, e. g., systems analysis and systems engineering, which together with OR have been called “hard systems thinking” (p. 127 in [31]). Also, more recent developments such as mathematical complexity and network theories, agent-based modeling and most versions of game theory can be classified as hard systems approaches.

The respective approaches have not altogether been excluded from fertile contacts with the interpretivist strand of inquiry. In principle, all of them can be considered as instruments for supporting discourses about different interpretations of an organizational reality or alternative futures studied in concrete cases. In our time, most applications of the VSM, for example, are constructivist in nature. To put it in a nutshell, these applications are (usually collective) constructions of a (new) reality, in which observation and interpretation play a crucial part. In this process, the actors involved make sense of the system under study, i. e., the organization in focus, by mapping it on the VSM. At the same time they bring forth “multiple realities rather than striving for a fit with one reality” (p. 299 in [29]).

The second group of methodologies is part of the interpretive strand:

- *Interactive Planning.* IP is a methodology, designed by Russell Ackoff [1], and developed further by Jamshid

Gharajedaghi [28], for the purpose of dealing with “messes” and enabling actors to design their desired futures, as well as to bring them about. It is grounded in theoretical work on purposeful systems, reverts to the principles of continuous, participative and holistic planning, and centers on the idea of an “idealized design.”

- *Soft Systems Methodology*. SSM is a heuristic designed by Peter Checkland [13,14] for dealing with complex situations. Checkland suggests a process of inquiry constituted by two aspects: A conceptual one, which is logic based, and a sociopolitical one, which is concerned with the cultural feasibility, desirability and implementation of change.
- *Critical Systems Heuristics*. CSH is a methodology, which Werner Ulrich [67,68] proposed for the purpose of scientifically informing planning and design in order to lead to an improvement in the human condition. The process aims at uncovering the interests that the system under study serves. The legitimacy and expertise of actors, and particularly the impacts of decisions and behaviors of the system on others – the “affected” – are elicited by means of a set of boundary questions. CSH can be seen as part of a wider movement known as the “Emancipatory Systems Approach” which embraces, e. g., Freire’s Critical Pedagogy, Interpretive Systemology, and Community OR (see pp. 291ff in [31]).

All three of these methodologies (IP, SSM, and CSH) are positioned in the interpretive tradition. Other methodologies and concepts which can be subsumed under the interpretive systems approach are, e. g., Warfield’s science of generic design, Churchman’s social system design, Senge’s soft systems thinking, Mason and Mitroff’s strategic assumptions surfacing and testing (SAST), Eden and Ackermann’s strategic options in development and analysis (SODA), and other methodologies of soft operational research (for details, see pp. 211ff in [31]). The interpretive methodologies were designed to deal with qualitative aspects in the analysis and design of complex systems, emphasizing the communicational, social, political and ethical dimensions of problem solving. Several authors mention explicitly that they do not preclude the use of quantitative techniques or include such techniques in their repertoire (e. g., the biocyberneticist Frederic Vester).

In an advanced understanding of system dynamics both of these traditions—positivist and interpretivist—are synthesized. The adherents of SD conceive of model building and validation as a semi-formal, relativistic, holistic social process. Validity is understood as usefulness or fit-

ness in relation to the purpose of the model, and validation as an elaborate set of procedures – including logico-structural, heuristic, algorithmic, statistical, and also discursive components – by which the quality of and the confidence in a model are gradually improved (see [4,5,59]).

### System Dynamics – Its Features, Strengths and Limitations

The features, strengths and limitations of the SD methodology are a consequence of its specific characteristics. In the context of the multiple theories and methodologies of the systems movement, some of the distinctive features of SD are (for an overview, see [52], pp. 142ff in [31]):

- *Feedback as conceptual basis*. SD model systems are high-order, multiple-loop networks of closed loops of information. Concomitantly, an interest in non-linearities, long-term patterns and internal structure rather than external disturbances is characteristic of SD (p. 31 in [40]). However, SD models are not “closed systems”, as sometimes is claimed, in the sense that (a) flows can originate from outside the system’s boundaries, (b) representations of exogenous factors or systems can be incorporated into any model as parameters or special modules, and (c) new information can be accommodated via changes to a model. In other words, the SD view hinges on a view of systems which are closed in a causal sense but not materially (p. 297 in [52]).
- *Focus on internally generated dynamics*. SD models are conceived as closed systems. The interest of users is in the dynamics generated inside those systems. Given the nature of closed feedback loops and the fact that delays occur within them, the dynamic behavior of these systems is essentially non-linear.
- *Emphasis on understanding*. For system dynamicists the understanding of the dynamics of a system is the first goal to be achieved by means of modeling and simulation. Conceptually, they try to understand events as embedded in patterns of behavior, which in turn are generated by underlying structures. Such understanding is enabled by SD as it “shows how present policies lead to future consequences” (Sect. VIII in [23]). Thereby, the feedback loops are “a major source of puzzling behavior and policy difficulties” (p. 300 in [52]). SD models purport to test mental models, hone intuition and improve learning (see [65]).
- *High degree of operationality*. SD relies on formal modeling. This fosters disciplined thinking; assumptions, underlying equations and quantifications must be clarified. Feedback loops and delays are visualized and formalized; therewith the causal logic inherent in a model



is made more transparent and discussable than in most other methodologies [53]. Also, a high level of realism in the models can be achieved. SD is therefore apt to support decision-making processes effectively.

- *Far-reaching requirements (and possibilities) for the combination of qualitative and quantitative aspects of modeling and simulation.* This is a consequence of the emphasis on understanding. The focus is not on point-precise prediction, but on the generation of insights into the patterns generated by the systems under study.
- *High level of generality and scale robustness.* The representation of dynamic systems in terms of stocks and flows is a generic form, which is adequate for a wide spectrum of potential applications. This spectrum is both broad as to the potential subjects under study, and deep as to the possible degrees of resolution and detail [38]. In addition, the SD methodology enables one to deal with large numbers of variables within multiple interacting feedback loops (p. 9 in [22]). SD has been applied to the most diverse subject areas, e.g., global modeling, environmental issues, social and economic policy, corporate and public management, regional planning, medicine, psychology and education in mathematics, physics and biology.

The features of SD just sketched out result in both strengths and limitations. We start with the strengths.

### Strengths of SD

1. Its *specific modeling approach* makes SD particularly helpful in gaining insights into the patterns exhibited by dynamic systems, as well as the structures underlying them. Closed-loop modeling has been found most useful in fostering understanding of the dynamic functioning of complex systems. Such understanding is especially facilitated by the principle of modeling the systems or issues under study in a continuous mode and at rather high aggregation levels [20,38]. With the help of relatively small but insightful models, and by means of sensitivity analyses as well as optimization heuristics incorporated in the application software packages, decision-spaces can be thoroughly explored. Vulnerabilities and the consequences of different system designs can be examined with relative ease.
2. The *generality of the methodology* and its power to crystallize operational thinking in realistic models have triggered applications in the most varied contexts. Easy-to-use software and the features of screen-driven modeling via graphic user interfaces provide a strong lever for collaborative model-building in teams (cf. [2,69]).

3. Another strong point is the *momentum of the SD movement*. Due to the strengths commented above this point, the community of users has grown steadily, being probably the largest community within the systems movement. Lane (p. 484 in [36]) has termed SD “one of the most widely used systems approaches in the world.”
4. Its specific features make SD an exceptionally effective tool for *conveying systemic thinking* to anybody. Therefore, it also has an outstanding track-record of classroom applications for which “learner-directed learning” [24] or “learner-centered learning” is advocated [25,26]. Pertinent audiences range from schoolchildren at the levels of secondary and primary schools to managers and scientists.

Given these strengths, the community of users has not only grown significantly, but has also transcended disciplinary boundaries, ranging from the formal and natural sciences to the humanities, and covering multiple uses from theory building and education to the tackling of real-world problems at almost any conceivable level. Applications to organizational, societal and ecological issues have seen a particularly strong growth. This feeds back on the availability and growth of the knowledge upon which the individual modeler can draw.

The flip side of most of the strengths outlined here embodies the limitations of SD; we concentrate on those which can be relevant to a possible complementarity of SD with other systems methodologies.

### Limitations of SD

1. The main point here is that SD does not provide a framework or methodology for the *diagnosis and design* of organizational structures in the sense of inter-relationships among organizational actors. This makes SD susceptible to completion from without – a completion which organizational cybernetics (OC), and the VSM in particular, but also living system theory (LST), especially can provide. The choice falls on these two approaches because of their strong heuristic power and their complementary strengths in relation to SD (cf. [57,61]).
2. Another limitation of SD is related to the *absorption of variety* (complexity) by an organization. *Variety* is a technical term for *complexity*, which denotes a (high) number of potential states or behaviors of a system (based on [3,8]). SD offers an approach to the handling of variety which allows modeling at different scales of a problem or system [47]. It focuses on the identification, at a certain resolution level or possibly several resolution levels, of the main stock variables which will be



affected by the respective flows. These, in turn, will be influenced by parameters and auxiliary variables. This approach, even though it enables thinking and modeling at different scales, does not provide a formal procedure for an organization to cope with the external complexity it faces, namely, for designing a structure which can absorb that complexity. In contrast, OC and LST offer elaborate models to enable the absorption of variety, in the case of the VSM based explicitly on Ashby's *Law of Requisite Variety*. It says "Only variety can destroy variety", which implies that the varieties of two interacting systems must be in balance, if stability is to be achieved [3]. The VSM has two salient features in this respect. Firstly, it helps design an organizational unit for viability, by enabling it to attenuate the complexity of its environment, and also to enhance its eigenvariety, so that the two are in balance. The term *variety engineering* has been used in this context [9]. Secondly, the recursive structure of the VSM ensures that an organization with several levels will develop sufficient eigenvariety along the fronts on which the complexity it faces unfolds. Similarly, LST offers the conditions for social systems to survive, by maintaining thermodynamically highly improbable energy states via continuous interaction with their environments. The difference between the two approaches is that the VSM functions more in the strategic and informational domains, while the LST model essentially focuses on the operational domain. In sum, both can make a strong contribution related to coping with the external complexity faced by organizations, and therefore can deliver a strong complement to SD.

3. Finally, the design of *modeling processes* confronts SD with specific challenges. The original SD methodology of modeling and simulation was to a large extent functionally and technically oriented. This made it strong in the domain of logical analysis, while the socio-cultural and political dimensions of the modeling process were, if not completely out of consideration, at least not a significant concern in methodological developments. The SD community – also under the influence of the soft systems approaches – has become aware of this limitation and has worked on incorporating features of the social sciences into its repertoire. The following examples, which document this effort to close the gap, stand for many. Extensive work on group model building has been achieved, which explores the potential of collaborative model building [69]. A new schema for the modeling process has been proposed, which complements logic-based analysis by cultural analysis [37]. The social dimension of system dynamics-based modeling has be-

come subject to intensive discussion ([77]; and other contributions to the special issue of *Systems Research and Behavioral Science*, Vol. 51, No. 4, 2006). Finally, in relation to consultancy methodology, modeling has been framed as a learning process [34] and as second-order intervention [60].

As has been shown, there is a need to complement classical SD with other methodologies, when issues are at stake which it cannot handle by itself. VSM and LST are excellent choices when issues of organizational diagnosis or design are to be tackled.

The limitations addressed here call attention to other methodologies which exhibit certain features that traditionally were not incorporated, or at least not explicit, in SD methodology. One aspect concerns the features that explicitly address the subjectivity of purposes and meanings ascribed to systems. In this context, support for problem formulation, model construction and strategy design by individuals on the one hand and groups on the other are relevant issues. Also, techniques for an enhancement of creativity (e.g., the generation and the reframing of options) in both individuals and groups are a matter of concern. Two further aspects relate to methodological arrangements for coping with the specific issues of negotiation and alignment in pluralist and coercive settings.

As far as the modeling processes are concerned, group model building has proven to be a valuable complement to pure modeling and simulation. However, there are other systems methodologies which should be considered as potentially apt to enrich SD analysis, namely the soft approaches commented upon earlier, e.g., interactive planning, soft system methodology and critical system heuristics.

On the other hand, SD can be a powerful complement to other methodologies which are more abstract or more static in nature. This potential refers essentially to all systems approaches which stand in the interpretive ("soft") tradition, but also to approaches which stand in the positivist traditions, such as the VSM and LST. These should revert to the support of SD in the event that tradeoffs between different goals must be handled, or if implications of long-term decisions on short-term outcomes (and vice versa) have to be ascertained, and whenever contingencies or vulnerabilities must be assessed.

### Actual and Potential Relationships

It should be clear by now that the systems movement has bred a number of theories and methodologies, none of which can be considered all-embracing or complete. All of

them have their strengths and weaknesses, and their specific potentials and limitations.

Since Burrell and Morgan [12] adverted to incommensurability between different paradigms of social theory, several authors have acknowledged or even advocated methodological complementarity. They argue that there is a potential complementarity between different methods, and, one may add, models, even if they come from distinct paradigms. Among these authors are, e.g., Brocklesby [11], Jackson [30], Midgley [43], Mingers [45], Schwaninger [55] and Yolles [83]. These authors have opened up a new perspective in comparison with the non-complementaristic state-of-the-art.

In the past, the different methodologies have led to the formation of their own traditions and “schools,” with boundaries across which not much dialogue has evolved. The methodologies have kept their protagonists busy testing them and developing them further. Also, the differences between different language games and epistemological traditions have often suggested incommensurability, and therewith have impaired communication. Prejudices and a lack of knowledge of the respective other side have accentuated this problem: Typically, “hard” systems scientists are suspicious of “soft” systems scientists. For example, many members of the OR community, not unlike orthodox quantitatively oriented economists, adhere to the opinion that “SD is too soft.” On the other hand the protagonists of “soft” systems approaches, even though many of them have adopted feedback diagrams (causal loop diagrams) for the sake of visualization, are all too often convinced that “SD is too hard.” Both of these judgments indicate a lack of knowledge, in particular of the SD validation and testing methods available, on the one hand, and the technical advancements achieved in modeling and simulation, on the other (see [5,59,66]).

In principle, both approaches are complementary. The qualitative view can enrich quantitative models, and it is connected to their philosophical, ethical and esthetical foundations. However, qualitative reasoning tends to be misleading if applied to causal network structures without being complemented by formalization and quantification of relationships and variables. Furthermore, the quantitative simulation fosters insights into qualitative patterns and principles. It is thus a most valuable device for validating and honing the intuition of decision makers, via corroboration and falsification.

Proposals that advocate mutual learning between the different “schools” have been formulated inside the SD community (e.g., [35]). The International System Dynamics Conference of 1994 in Stirling, held under the banner of “Transcending the Boundaries,” was dedicated to the dia-

logue between different streams of the systems movement.

Also, from the 1990s onwards, there were vigorous efforts to deal with methodological challenges, which traditionally had not been an important matter of scientific interest within the SD community. Some of the progress made in these areas is documented in a special edition of *Systems Research and Behavioral Science* (Vol. 21, No. 4, July-August 2004). The main point is that much of the available potential is based on the complementarity, not the mutual exclusiveness, of the different systems approaches.

In the future, much can be gained from leveraging these complementarities. Here are two examples of methodological developments in this direction, which appear to be achievable and potentially fertile: The enhancement of qualitative components in “soft” systems methodologies in the process of knowledge elicitation and model building (cf. [69]), and the combination of cybernetics-based organizational design with SD-based modeling and simulation (cf. [61]). Potential complementarities exist not only across the qualities – quantities boundary, but also within each one of the domains. For example, with the help of advanced software, SD modeling (“top-down”) and agent-based modeling (“bottom-up”) can be used in combination.

From a meta-methodological stance, generalist frameworks have been elaborated which contain blueprints for combining different methodologies where this is indicated. Two examples are:

- *Total systems intervention* (TSI) is a framework proposed by Flood and Jackson [19], which furnishes a number of heuristic schemes and principles for the purpose of selecting and combining systems methods/methodologies in a customized way, according to the issue to be tackled. SD is among the recommended “tools”.
- *Integrative systems methodology* (ISM) is a heuristic for providing actors in organizations with requisite variety, developed by Schwaninger [55,56]. It advocates (a) dealing with both content- and context-related issues during the process, and (b) placing a stronger emphasis on the validation of qualitative and quantitative models as well as strategies, in both dimensions of the content of the issue under study and the organizational context into which that issue is embedded. For this purpose, the tools of SD (to model content) and organizational cybernetics – the VSM (to model context) – are cogently integrated.

These are only two examples. In principle, SD could make an important contribution in the context of most of

the methodological frameworks, far beyond the extent to which this has been the case. Systems methodologists and practitioners can potentially benefit enormously from including SD methodology in their repertoires.

## Outlook

There have recently been calls for an eclectic “mixing and matching” of methodologies. In light of the epistemological tendencies of our time towards radical relativism, it is necessary to warn against taking a course in which “anything goes”. It is most important to emphasize that the desirable methodological progress can only be achieved on the grounds of scientific rigor. This postulate of “rigor” is not to be confused with an encouragement of “rigidity.” The necessary methodological principles advocated here are disciplined thinking, a permanent quest for better models (that is, thorough validation), and the highest achievable levels of transparency in the formalizations as well as of the underlying assumptions and sources used. Scientific rigor, in this context, also implies that combinations of methodologies reach beyond merely eclectic additions from different methodologies, so that genuine integration towards better adequacy to the issues at hand is achieved.

The contribution of system dynamics can come in the realms of the following:

- Fostering disciplined thinking
- Understanding dynamic behaviors of systems and the structures that generate them
- Exploring paths into the future and the concrete implications of decisions
- Assessing strategies as to their robustness and vulnerabilities, in ways precluded by other, more philosophical, and generally “soft” systems approaches

These latter streams can contribute to reflecting and tackling the meaning- and value-laden dimensions of complex human, social and ecological systems. Some of their features should and can be combined synergistically with system dynamics, particularly by being incorporated into the repertoires of system dynamicists. From the reverse perspective, incorporating system dynamics as a standard tool will be of great benefit for the broad methodological frameworks. Model formalization and dynamic simulation may even be considered necessary components for the study of the concrete dynamics of complex systems.

Finally, there are also many developments in the “hard”, i.e., mathematics-, statistics-, logic-, and informatics-based methods and technologies, which are apt to enrich the system dynamics methodology, namely in

terms of modeling and decision support. For example, the constantly evolving techniques of time-series analysis, filtering, neural networks and control theory can improve the design of system-dynamics-based systems of (self-)control. Also, a bridge across the divide between the top-down modeling approach of SD and the bottom-up approach of agent-based modeling appears to be feasible. Furthermore, a promising perspective for the design of genuinely “intelligent organizations” emerges if one combines SD with advanced database-management, cooperative model building software, and the qualitative features of the “soft” systems methodologies.

The approaches of integrating complementary methodologies outlined in this contribution definitely mark a new phase in the history of the systems movement.

## Appendix

### Milestones in the Evolution of the Systems Approach in General and System Dynamics in Particular

The table gives an overview of the systems movement’s evolution, as shown in its main literature; and that overview is not exhaustive.

### Systems Approaches – An Overview

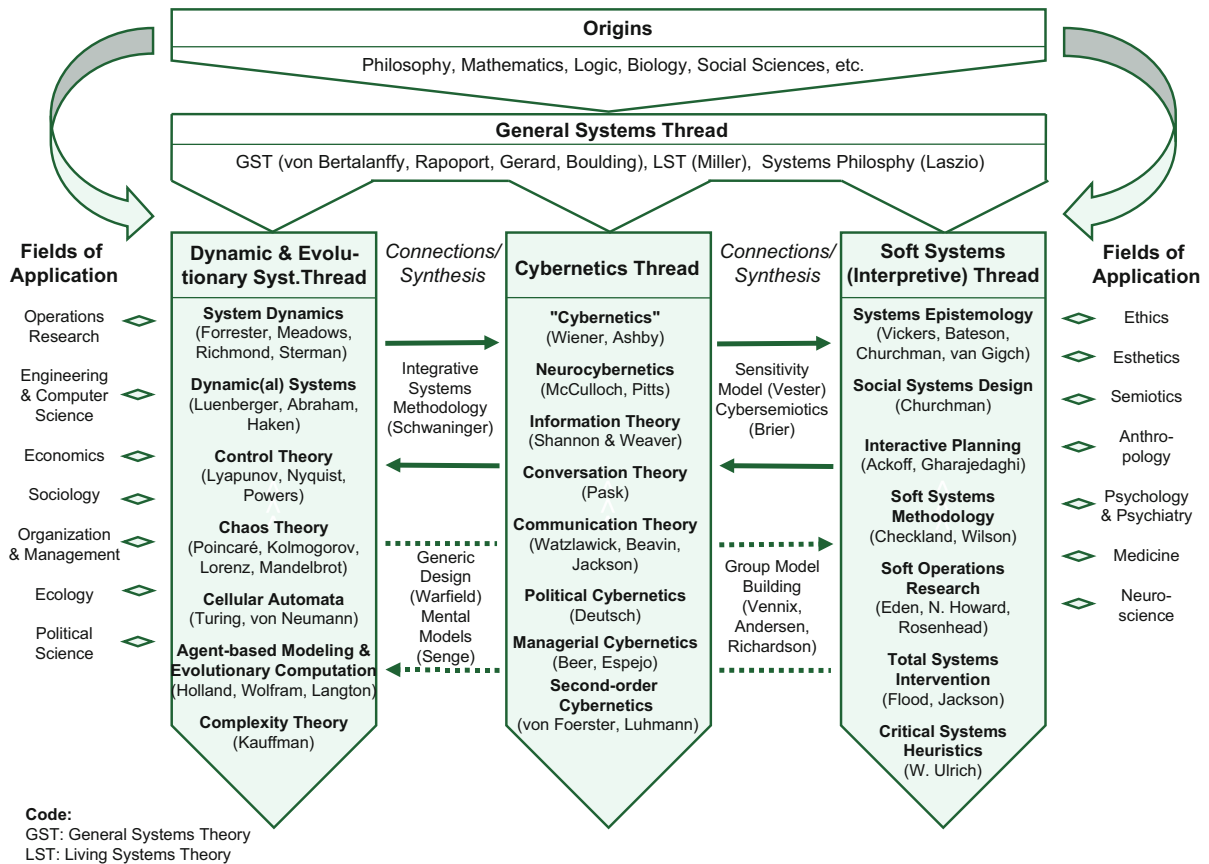
Note: This diagram shows three streams of the systems approach in the context of their antecedents. The general systems thread has its origins in philosophical roots from antiquity: The term *system* derives from the old Greek *σύστημα* (*systema*), while, *cybernetics* stems from the Greek *κυβερνήτης* (*kybernētēs*). The arrows between the threads stand for interrelationships and efforts to synthesize the connected approaches. For example, integrated systems methodology is an integrative attempt to leverage the complementarities of system dynamics and organizational cybernetics. Enumerated to the left and right of the scheme are the fields of application. The big arrows in the upper region of the diagram indicate that the roots of the systems approach continue influencing the different threads and the fields of application even if the path via general systems theory is not pursued.

The diagram is not a complete representation, but the result of an attempt to map the major threads of the systems movement and some of their interrelations. Hence, the schema does not cover all schools or protagonists of the movement. Why does the diagram show a dynamic and evolutionary systems thread and a cybernetics thread, if cybernetics is about dynamic systems? The latter embraces all the approaches that are explicitly grounded in cybernetics. The former relates to all other approaches

System Dynamics in the Evolution of the Systems Approach, Table 1

Milestones in the evolution of the systems approach in general and system dynamics in particular

|                                                   |                                                                                        |            |
|---------------------------------------------------|----------------------------------------------------------------------------------------|------------|
| <b>Foundations of general system theory</b>       |                                                                                        |            |
| Von Bertalanffy                                   | Zu einer allgemeinen Systemlehre                                                       | 1945       |
|                                                   | An Outline of General System Theory                                                    | 1950       |
|                                                   | General System Theory                                                                  | 1968       |
| Bertalanffy, Boulding, Gerard, Rapoport           | Foundation of the Society for General Systems Research                                 | 1953       |
| Klir                                              | An Approach to General System Theory                                                   | 1968       |
| Simon                                             | The Sciences of the Artificial                                                         | 1969       |
| Pichler                                           | Mathematische Systemtheorie                                                            | 1975       |
| Miller                                            | Living Systems                                                                         | 1978       |
| Mesarovic & Takahara                              | Abstract Systems Theory                                                                | 1985       |
| Rapoport                                          | General System Theory                                                                  | 1986       |
| <b>Foundations of cybernetics</b>                 |                                                                                        |            |
| Macy Conferences<br>(Josiah Macy, Jr. Foundation) | Cybernetics. Circular Causal, and Feedback Mechanisms in Biological and Social Systems | 1946–1951  |
| Wiener                                            | Cybernetics or Control and Communication in the Animal and in the Machine              | 1948       |
| Ashby                                             | An Introduction to Cybernetics                                                         | 1956       |
| Pask                                              | An Approach to Cybernetics                                                             | 1961       |
| Von Foerster, Zopf                                | Principles of Self-Organization                                                        | 1962       |
| McCulloch                                         | Embodiments of Mind                                                                    | 1965       |
| <b>Foundations of organizational cybernetics</b>  |                                                                                        |            |
| Beer                                              | Cybernetics and Management                                                             | 1959       |
|                                                   | Towards the Cybernetic Factory                                                         | 1962       |
|                                                   | Decision and Control                                                                   | 1966       |
|                                                   | Brain of the Firm                                                                      | 1972       |
| Von Foerster                                      | Cybernetics of Cybernetics                                                             | 1974       |
| <b>Foundations of system dynamics</b>             |                                                                                        |            |
| Forrester                                         | Industrial Dynamics                                                                    | 1961       |
|                                                   | Principles of Systems                                                                  | 1968       |
|                                                   | Urban Dynamics                                                                         | 1969       |
|                                                   | World Dynamics                                                                         | 1971       |
| Meadows et al.                                    | Limits to Growth                                                                       | 1972       |
| Richardson                                        | Feedback Thought in Social Science and Systems Theory                                  | 1991       |
| <b>Systems methodology</b>                        |                                                                                        |            |
| Churchman                                         | Challenge to Reason                                                                    | 1968       |
|                                                   | The Systems Approach                                                                   | 1968       |
| Vester & von Hesler                               | Sensitivitätsmodell                                                                    | 1980       |
| Checkland                                         | Systems Thinking, Systems Practice                                                     | 1981       |
| Ackoff                                            | Creating the Corporate Future                                                          | 1981       |
| Ulrich                                            | Critical Heuristics of Social Planning                                                 | 1983       |
| Warfield                                          | A Science of Generic Design                                                            | 1994       |
| Schwaninger                                       | Integrative Systems Methodology                                                        | 1997       |
| Gharajedaghi                                      | Systems Thinking                                                                       | 1999       |
| Sabelli                                           | Bios – A Study of Creation                                                             | 2005       |
| <b>Selected recent works in system dynamics</b>   |                                                                                        |            |
| Senge                                             | The Fifth Discipline                                                                   | 1990       |
| Barlas & Carpenter                                | Model Validity                                                                         | 1990       |
| Vennix                                            | Group Model Building                                                                   | 1996       |
| Lane & Oliva                                      | Synthesis of System Dynamics and Soft Systems Methodology                              | 1998       |
| Sterman                                           | Business Dynamics                                                                      | 2000       |
| Warren                                            | Strategy Dynamics                                                                      | 2002, 2008 |
| Wolstenholme                                      | Archetypal Structures                                                                  | 2003       |
| Morecroft                                         | Strategic Modelling                                                                    | 2007       |
| Schwaninger & Grösser                             | Theory-building with System Dynamics & Model Validation                                | 2008, 2009 |



System Dynamics in the Evolution of the Systems Approach, Figure 1

concerned with dynamic or evolutionary systems. The simplification made it necessary to somewhat curtail logical perfection for the sake of conveying a synoptic view of the different systems approaches, in a language that uses the categories common in current scientific and professional discourse. Overlaps exist, e. g., between dynamic systems and chaos theory, cellular automata and agent-based modeling.

## Bibliography

### Primary Literature

- Ackoff RL (1981) Creating the Corporate Future. Wiley, New York
- Andersen DF, Richardson GP (1997) Scripts for Group Model Building. Syst Dyn Rev 13(2):107–129
- Ashby WR (1956) An Introduction to Cybernetics. Chapman & Hall, London
- Barlas Y (1996) Formal aspects of model validity and validation in system dynamics. Syst Dyn Rev 12(3):183–210
- Barlas Y, Carpenter S (1990) Philosophical roots of model validity: Two paradigms. Syst Dyn Rev 6(2):148–166
- Beer S (1959) Cybernetics and Management. English Universities Press, London
- Beer S (1994) Towards the Cybernetic Factory. In: Harnden R, Leonard A (eds) How Many Grapes Went into the Wine. Stafford Beer on the Art and Science of Holistic Management. Wiley, Chichester, pp 163–225 (reprint, originally published in 1962)
- Beer S (1966) Decision and Control. Wiley, Chichester
- Beer S (1979) The Heart of Enterprise. Wiley, Chichester
- Beer S (1981) Brain of the Firm, 2nd edn. Wiley, Chichester
- Brocklesby J (1993) Methodological complementarity or separate paradigm development – Examining the options for enhanced operational research. Aust J Manag 18(2):133–157
- Burrell G, Morgan G (1979) Sociological Paradigms and Organisational Analysis. Hants, Gower
- Checkland PB (1981) Systems Thinking, Systems Practice. Wiley, Chichester
- Checkland PB, Poulter J (2006) Learning for Action: A Short Definitive Account of Soft Systems Methodology, and its Use Practitioners, Teachers and Students. Wiley, Chichester
- Churchman CW (1968) Challenge to Reason. McGraw-Hill, New York
- Churchman CW (1968) The Systems Approach. Delacorte Press, New York
- Churchman CW (1979) The Systems Approach and its Enemies. Basic Books, New York
- Encyclopedia of Life Support Systems (2002) published under: <http://www.eolss.net/>



19. Flood RL, Jackson MC (1991) *Creative Problem Solving. Total Systems Intervention*. Wiley, Chichester
20. Forrester JW (1961) *Industrial Dynamics*. MIT Press, Cambridge
21. Forrester JW (1968) *Principles of Systems*. MIT Press, Cambridge
22. Forrester JW (1969) *Urban Dynamics*. MIT Press, Cambridge
23. Forrester JW (1971) *World Dynamics*. Pegasus Communications, Waltham
24. Forrester JW (1993) *System Dynamics and the Lessons of 35 Years*. In: DeGreene KB (ed) *Systems-Based Approach to Policy Making*. Kluwer, Boston
25. Forrester JW (1993) *System Dynamics as an Organizing Framework for Pre-college Education*. *Syst Dyn Rev* 9(2):183–194
26. Forrester JW (1997) *System Dynamics and K-12 Teachers. A Lecture at the University of Virginia School of Education, Massachusetts Institute of Technology*. System Dynamics Group Paper D-4665–4
27. François C (2004) *International Encyclopedia of Systems and Cybernetics*, 2nd edn. Saur, München
28. Gharajedaghi J (1999) *Systems Thinking. Managing Chaos and Complexity*. Butterworth-Heinemann, Boston
29. Harnden RJ (1989) *Technology for Enabling: The Implications for Management Science of a Hermeneutics of Distinction*. The University of Aston, Birmingham
30. Jackson MC (1991) *Systems Methodology for the Management Sciences*. Plenum Press, New York
31. Jackson MC (2000) *Systems Approaches to Management*. Kluwer Academic/Plenum, New York
32. Kauffman SA (1993) *The Origins of Order. Self-Organization and Selection in Evolution*. Oxford University Press, New York
33. Klir GJ (1969) *An Approach to General Systems Theory*. Nostrand, New York
34. Lane DC (1994) *Modeling as Learning: A Consultancy Methodology for Enhancing Learning in Management in Management Teams*. In: Morecroft J, Sterman JD (eds) *Modeling for Learning Organizations*. Productivity Press, Portland, pp 205–240
35. Lane DC (1994) *With a little help from our friends: How system dynamics and soft OR can learn from each other*. *Syst Dyn Rev* 10(2–3):101–134
36. Lane DC (2006) *IFORS' Operational Research Hall of Fame*. Jay Wright Forrester. *Int Trans Oper Res* 13:483–492
37. Lane DC, Oliva R (1998) *The Greater Whole: towards a synthesis of system dynamics and soft system methodology*. *Eur J Oper Res* 107(1):214–235
38. La Roche U, Simon M (2000) *Geschäftsprozesse simulieren: flexibel und zielorientiert führen mit Fließmodellen*. Orell Füssli, Zürich
39. McCulloch WS (1965) *Embodiments of Mind*. MIT Press, Cambridge
40. Meadows DH (1980) *The Unavoidable A Priori*. In: Randers J (ed) *Elements of the System Dynamics Method*. MIT Press, Cambridge, pp 23–57
41. Meadows DH, Meadows DL, Randers J, Behrens III WW (1972) *Limits to Growth*. Universe Books, New York
42. Mesarovic MD, Takahara Y (1985) *Abstract Systems Theory*. Springer, Berlin
43. Midgley G (2000) *Systemic Intervention. Philosophy, Methodology, and Practice*. Kluwer, New York
44. Miller JG (1978) *Living Systems*. McGraw-Hill, New York
45. Mingers J (1997) *Multi-paradigm Multimethodology*. In: Mingers J, Gill A (eds) *Multimethodology*. Wiley, Chichester
46. Morecroft J (2007) *Strategic Modelling and Business Dynamics: a Feedback Systems Approach*. Wiley, Chichester
47. Odum HT, Odum EC (2000) *Modeling for all Scales: An Introduction to System Simulation*. Academic Press, San Diego
48. Pask G (1961) *An Approach to Cybernetics*. Hutchinson, London
49. Pichler F (1975) *Mathematische Systemtheorie*. de Gruyter, Berlin
50. Rapoport A (1953) *Operational philosophy: Integrating Knowledge and Action*. Harper, New York
51. Rapoport A (1986) *General System Theory. Essential Concepts and Applications*. Abacus Press, Turnbridge Wells
52. Richardson GP (1999) *Feedback Thought in Social Science and Systems Theory*. Pegasus Communications, Waltham (Originally published in 1991)
53. Richmond B (1997) *The "Thinking" in systems thinking: How can we make it easier to master?* *Syst Think* 8(2):1–5
54. Sabelli H (2005) *Bios: a Study of Creation*. World Scientific, Hackensack
55. Schwaninger M (1997) *Integrative systems methodology: Heuristic for requisite variety*. *Int Trans Oper Res* 4(4):109–123
56. Schwaninger M (2004) *Methodologies in conflict: Achieving synergies between system dynamics and organizational cybernetics*. *Syst Res Behav Sci* 21(4):1–21
57. Schwaninger M (2006) *Theories of viability. A comparison*. *Syst Res Behav Sci* 23:337–347
58. Schwaninger M, Groesser S (2008) *System dynamics as model-based theory building*. *Syst Res Behav Sci* 25:1–19
59. Schwaninger M, Groesser S (2009) *Model Validation: The Quest for Quality in System Dynamics Modeling*. *Encyclopaedia of Complexity and Systems Science*. Springer, New York
60. Schwaninger M, Janovjak M, Ambroz K (2006) *Second-order intervention: Enhancing organizational competence and performance*. *Syst Res Behav Sci* 23:529–545
61. Schwaninger M, Pérez Ríos J (2008) *System dynamics and cybernetics: A synergetic pair*. *Syst Dyn Rev* 24(2):145–174
62. Senge PM (1990) *The Fifth Discipline. The Art and Practice of the Learning Organization*. Doubleday, New York
63. Shapiro M, Mandel T, Schwaninger M et al (1996) *The Primer Toolbox*. International Society for the Systems Sciences, <http://www.iss.org/primer/toolbox.htm>
64. Simon HA (1969) *The Sciences of the Artificial*. MIT Press, Cambridge
65. Sterman JD (1994) *Learning in and about complex systems*. *Syst Dyn Rev* 10(2–3):291–330
66. Sterman JD (2000) *Business Dynamics. Systems Thinking and Modeling for a Complex World*. Irwin/McGraw-Hill, Boston
67. Ulrich W (1983) *Critical Heuristics of Social Planning*. Haupt, Bern
68. Ulrich W (1996) *A Primer to Critical Systems Heuristics for Action Researchers*. The Centre of Systems Studies, University of Hull, Hull
69. Vennix JAM (1996) *Group Model Building. Facilitating Team Learning Using System Dynamics*. Wiley, Chichester
70. Vester F, Von Hesler A (1980) *Sensitivitätsmodell. Regionale Planungsgemeinschaft Untermain, Frankfurt am Main*
71. Von Bertalanffy L (1949) *Zu einer allgemeinen Systemlehre*. *Bl Dtsch Philos* 18(3/4) (Excerpts: in *Biol Gen* 19(1):114–129 and in *General System Theory*, 1968, Chapter III)
72. Von Bertalanffy L (1950) *An outline of general system theory*. *Br J Philos Sci* 1:139–164

73. Von Bertalanffy L (1968) General System Theory. Braziller, New York
74. Von Foerster H (1984) Observing Systems, 2nd edn. Intersystems Publications, Seaside
75. Von Foerster H (ed) (1995) Cybernetics of Cybernetics, 2nd edn. Future Systems, Minneapolis (Originally published in 1974)
76. Von Foerster H, Zopf GW (eds) (1962) Principles of Self-Organization. Pergamon Press, Oxford
77. Vriens D, Achtenbergh J (2006) The social dimension of system dynamics-based modelling. *Syst Res Behav Sci* 23(4): 553–563
78. Warfield JN (1994) A Science of Generic Design: Managing Complexity through Systems Design, 2nd edn. Iowa State University Press, Ames
79. Warren K (2002) Competitive Strategy Dynamics. Wiley, Chichester
80. Warren K (2008) Strategic Management Dynamics. Wiley, Chichester
81. Wiener N (1948) Cybernetics: Control and Communication in the Animal and in the Machine. MIT Press, Cambridge
82. Wolstenholme E (2003) Towards the definition and use of a core set of archetypal structures in system dynamics. *Syst Dyn Rev* 19(1):7–26
83. Yolles MA (1998) Cybernetic exploration of methodological complement. *Kybern* 27(5):527–542

### Books and Reviews

- Jackson MC (2003) Systems Thinking: Creative Holism for Managers. Wiley, Chichester
- Klir GJ (2001) Facets of Systems Science, 2nd edn. Kluwer Academic/Plenum, New York
- Midgley G (ed) (2003) Systems Thinking, vol 4. Sage, London
- Ragsdell G, Wilby J (eds) (2001) Understanding Complexity. Kluwer Academic/Plenum, New York
- Richardson GP (ed) (1996) Modelling for Management. Simulation in Support of Systems Thinking, 2 Volumes. Aldershot, Dartmouth
- Schwaninger M (2006) Intelligent Organizations. Powerful Model for Systemic Management. Springer, Berlin
- Van Gigh JP (2003) Metadecisions. Rehabilitating Epistemology. Kluwer Academic/Plenum, New York

## System Dynamics, Introduction to

BRIAN DANGERFIELD

Centre for OR & Applied Statistics, Salford Business School, University of Salford, Salford, UK

When Jay Wright Forrester published his first paper in 1958 he subtitled it “*a major breakthrough for decision-makers*”. At the time some thought this rather an exaggeration if not pompous. Now that 50 years of system dynamics (SD) has elapsed we can at least point to the achievements made and re-state continuing progress in

the pages of this section. Was it a ‘major breakthrough’? It certainly has the potential to raise the standards in evidence-based policy making to warrant this description and some startlingly good examples of such work will be mentioned here. But after 50 years perhaps one might expect more than has surfaced heretofore.

The key might be connected to the skills required to formulate good SD models – those which address a real-world problem with devastating simplicity and insight. It is deceptively easy to produce an SD model but there are subtleties involved in producing a really effective model for policy purposes. An uplift in modeling skills is something which a subset of the (now significant) amount of published material on SD is aimed at and this section will add to that corpus of work. In addition it will illustrate the extent to which SD applications have spread from its genesis in business to embrace health care, environmental, energy and climate issues, project management, some aspects of biological science and human physiology, governmental and public policy generally, economics (mainly macro), the diffusion of innovations and finally social and economic development. Other applications are being encountered as the power of the methodology is becoming appreciated. It has long since justified the change of title from *Industrial Dynamics* (1958) to *System Dynamics* (1970 onwards).

Richardson contributes an overview of the basics of SD modeling (see ► [System Dynamics, The Basic Elements of](#)). The underlying conceptual framework is that of the information feedback loop together with resource stocks and flows and an endogenous perspective on causation. The simplicity of the loop concept is apt to contribute to the apparent ease with which SD models can be created (along with the icon-based suites of SD software). But the novice reader should appreciate that it can take time to assimilate the modeling skills necessary to execute well an SD model-based application. Practice is essential and the references included will lead to further published material to assist the steep climb up the learning curve. So-called experts are still being confronted with the subtleties of SD modeling after years of involvement.

To place the SD methodology in context, the contribution by Schwaninger (see ► [System Dynamics in the Evolution of the Systems Approach](#)) profiles it alongside various others ‘systems’ based approaches which have emerged in the management and social sciences. Those professing to become experts in SD need to know about the other range of approaches which co-exist in the field of systems science. All these other methodologies have their own enthusiasts and this may even extend to the formation of societies with annual conferences. His Appendix B

shows a diagram of the different systems approaches and their interrelationships.

The foundations of the SD methodology can be characterized by certain philosophical issues. Olaya's text (see ► [System Dynamics Philosophical Background and Underpinnings](#)) defines a central one as presentationalism, associated with the notion of 'mental models'. A number of other philosophical issues which relate to SD are introduced, including those of positivism and social theory.

The practice of SD when applied to real-world applications essentially involves managerial learning and will often involve an interaction with client teams rather than one individual. How best to organize such structured approaches to participative model building is described by Rouwette and Vennix (see ► [Group Model Building](#)). Client participation is required for successful modeling.

If the promotion of learning and understanding is the primary *raison d'être* of SD, then achievement of this goal in an individual can be a significant accomplishment, especially if that person is the most senior in the client team. But there is a further goal to be pursued should the study fully reap the benefits of the SD methodology: How can we foster *organizational* learning? Maani tackles this head on (see ► [System Dynamics and Organizational Learning](#)). He defines the core capabilities of a learning organization and goes on to list the developing literature on organizational learning and, most importantly, how SD can aid the process through learning laboratories and microworlds.

Running an SD model creates a time-path of output behavior covering all the variables it is deemed necessary to include in the model. The various runs of the model are, most frequently, addressed in comparative fashion rather than taken in isolation. They can therefore be described as computer-based scenarios each of which charts a possible but not assured future. Georgantzas (see ► [Scenario-Driven Planning with System Dynamics](#)) describes environmental (traditional) scenario generation for which there is a considerable body of literature. But he emphasizes that successful strategy design involves the integration of three things: a knowledge of the business environment; the effects of unstated assumptions about change in the environment and strategy on performance; and finally the need to *compute* the effects on organizational performance. These three facets are accomplished by the process of SD modeling.

Thus far this introductory roadmap has covered all the background for contextualizing and creating an SD model. We now turn to various tasks associated with ex post modeling activities. Three such aspects are covered: model validation; analytical methods to explain behavior and determine dominant loops; and model optimization.

Schwaninger and Groesser (see ► [System Dynamics Modeling: Validation for Quality Assurance](#)) range over the various aspects of model validation, beginning with its epistemological foundations. In real-world modeling studies testing and validation is a sine qua non of the process. The range of tests made available and the attention given to the task of validation in the literature mark out SD as unique in the field of management science. Few other methodologies get near to the variety of tests which can be applied to an SD model. The authors consider the range of tests under three headings: model-related context; model structure; and model behavior.

Kampmann and Oliva deal with the behavioral analysis issue (see ► [System Dynamics, Analytical Methods for Structural Dominance Analysis in](#)). This activity tries to shed light on the model's dynamic behavior: Why does it behave as it does? What loop structures are responsible for the dominant behavior – and indeed shifts in that behavior where it occurs? In other words, they explore the link between system structure and dynamic behavior. Early methods used eigenvalue analysis but, since then, more sophisticated approaches have been put forward. A major advance will occur when one or more of these is refined enough to be included in an SD software package. This is likely to take some time although an improved user interface showing links glowing with differing degrees of intensity, reflecting their relative importance, is possible in the not-too-distant future.

Dangerfield describes the methods for improving model performance (see ► [System Dynamics Models, Optimization of](#)). The task can be categorized under two headings: calibration and policy optimization. The former relates to the determination of optimal parameter sets which deliver the best fit of the model to past time series data. Policy optimization on the other hand seeks to establish policies which deliver the 'best' performance against a suitable metric, such as minimum cost or maximum revenue. Using such an approach can accelerate the learning which comes from repeated runs of the model. Sadly, in the existing SD literature, there is scant evidence of its use in real-world studies.

The methodology of SD exists for no other reason than to offer a quantum leap in the standards of policy analysis. Therefore, any review must include a range through the landscape which defines areas of application. There are eight such areas covered in this section and the choice has been made in the knowledge that there are others which may also have been included and some new areas which are only just being opened up to the tools of SD modeling.

Business Strategy was the genesis of SD applications and rightly takes pride of place. This is the field in which

the most numerous SD applications occur. Lyneis (see ► [Business Policy and Strategy, System Dynamics Applications to](#)) concentrates on the process of how SD models are used in the task of strategy formulation. He goes on to consider the various drivers of business dynamics such as oscillations in supply chains and boom and bust life cycles. Detailed references are provided for a wide range of business application case studies.

Health care is consuming a higher share of GDP in many Western industrialized countries. This is due to the age profile of the population and advances in pharmacological and medical technologies. It is unsurprising that SD methods have been applied in tackling some of the most high-profile issues in health care and the relatively recent literature is testimony to the success of SD-based analysis. Indeed, it is arguable that some of the best modeling applications have surfaced in this sector. To do justice to the field of health care two contributions were solicited, in part because of the different funding systems which exist on either side of the Atlantic: Wolstenholme surveys the work done by UK and European authors (see ► [Health Care in the United Kingdom and Europe, System Dynamics Applications to](#)), whilst Hirsch and Homer concentrate on work published by US authors (see ► [Health Care in the United States, System Dynamics Applications to](#)).

Wolstenholme describes work carried out in the UK and Continental Europe but gives particular emphasis to three areas where models have been deployed. He starts with the problem of delayed hospital discharge which generates hospital capacity problems. Epidemiology is also reviewed, in particular research on the epidemiology of HIV/AIDS. Finally, recent work on mental health reform in the UK is described.

Hirsch and Homer note that the system in the USA is comparatively difficult to manage because of its free-market approach and relative lack of regulation. They concentrate on three main areas: disease epidemiology including heart disease and diabetes; substance abuse; and health care capacity and delivery.

Along with health care, the depletion of environmental resources and its effects has consumed many thousands of column inches in printed news media. SD has been employed in the pursuit of more compelling applications in this sector and the efforts go back to the well-known *Limits to Growth* study in 1971–72. Ford charts the most notable efforts which have emerged (see ► [System Dynamics Models of Environment, Energy and Climate Change](#)). He ranges over environmental resource problems in the western USA, models for greater understanding of climate change and global warming and concludes with studies in

energy, specifically two applications to the electric power industry.

The field of economics is one where SD has received a mostly hostile reception. The statistical economic modeling tool of econometrics has an extensive history and as a preferred modeling methodology seems hard to dislodge. However, there are an increasing number of heterodox economists who are prepared to embrace SD concepts and Radzicki (see ► [System Dynamics and Its Contribution to Economics and Economic Modeling](#)) describes the advances taking place. Whilst some of the literature embodies the translation of existing economic models into an SD format (which is a laudable objective) he calls for more economic dynamics models to be built from scratch embodying the best practice in SD modeling. Economic policy is too important to be informed by a single, seemingly unassailable, modeling methodology and it is to be hoped that in the future SD will become even more accepted as a viable tool for use in this field.

In a similar vein comes the contribution of Saeed (see ► [Dynamics of Income Distribution in a Market Economy: Possibilities for Poverty Alleviation](#)). He takes an economic modeling perspective and describes an SD model which explains resource allocation, production and entitlements in a market economy. Its purpose is to understand better how poverty might be reduced in the context of the redistribution of income. A comprehensive listing of the model is provided in an appendix.

The application of SD to public policy generally is dealt with by Andersen, Rich and MacDonald (see ► [Public Policy, System Dynamics Applications to](#)). They emphasize how public policy issues are complex, cross organizational boundaries, involve stakeholders with widely different perspectives and evolve over time, such that longer term results may be wholly different from short-term outcomes. Detail is provided for one public policy case involving the Governor's Office of Regulatory Assistance in New York State. They conclude with coverage of studies in a range of public domains such as defense, health care, education and the environment.

One area of SD application has brought the methodology into the legal arena. Disruption and delay in the execution of complex projects invariably finds two parties in dispute. Such disputes often center upon time delays and use of resources on projects – and what might have happened if things had been managed differently. SD models have been employed by parties to such disputes to attempt to justify the occurrence of these events. Howick, Ackermann, Eden and Williams (see ► [Delay and Disruption in Complex Projects](#)) report on how cognitive mapping, cause mapping and SD can be fused into what they de-



scribe as a cascade model building process. The result is a rigorous process for explaining why a project behaved in a certain way.

New products and processes are emerging at an ever-increasing rate in modern times. We need to understand the myriad mechanisms which are the basis for their rate of adoption. Milling and Maier range over various SD models which have been created to understand and improve the management of the diffusion of innovations (see ► [Diffusion of Innovations, System Dynamics Analysis of the](#)). From the often-cited Bass diffusion model (1969) the authors develop a series of additional features in a modular fashion. These features include competition, network externalities, dynamic pricing and research and development. They conclude by stressing how it is not possible to offer general recommendations for strategies in dynamic and complex environments; such recommendations can only be given in the context of the specific case under scrutiny.

## System Dynamics and Its Contribution to Economics and Economic Modeling

MICHAEL J. RADZICKI  
Worcester Polytechnic Institute, Worcester, USA

### Article Outline

- Glossary
- Definition of the Subject
- Introduction
- Types of Dynamic Simulation
- Translating Existing Economic Models into a System Dynamics Format
- Improving Existing Economic Models with System Dynamics
- Creating Economic Dynamics Models from Scratch
- Model Validity
- Controversies
- Future Directions
- Bibliography

### Glossary

**Stock** Stocks, which are sometimes referred to as “levels” or “states”, accumulate (i. e., sum up) the information or material that flows into and out of them. Stocks are thus responsible for decoupling flows, creating de-

lays, preserving system memory, and altering the time shape of flows.

**Flow** Flows of information or material enter and exit a system’s stocks and, in so doing, create a system’s dynamics. Stated differently, the net flow into or out of a stock is the stock’s rate of change. When human decision making is represented in a system dynamics model, it appears in the system’s flow equations. Mathematically, a system’s flow equations are ordinary differential equations and their format determines whether or not a system is linear or nonlinear.

**Feedback** Feedback is the transmission and return of information about the amount of information or material that has accumulated in a system’s stocks. When the return of this information reinforces a system’s behavior, the loop is said to be positive. Positive loops are responsible for the exponential growth of a system over time. Negative feedback loops represent goal seeking behavior in complex systems. When a negative loop detects a gap between the amount of information or material in a system’s stock and the desired amount of information or material, it initiates corrective action. If this corrective action is not significantly delayed, the system will smoothly adjust to its goal. If the corrective action is delayed, however, the system can overshoot or undershoot its goal and the system can oscillate.

**Full information maximum likelihood with optimal filtering** FIMLOF is a sophisticated technique for estimating the parameters of a system dynamics model, while simultaneously fitting its output to numerical data. Its intellectual origins can be traced to control engineering and the work of Fred Schwepe. David Peterson pioneered a method for adapting FIMLOF for use in system dynamics modeling.

### Definition of the Subject

System dynamics is a computer modeling method that has its intellectual origins in control engineering, management science, and digital computing. It was originally created as a tool to help managers better understand and control corporate systems. Today it is applied to problems in a wide variety of academic disciplines, including economics. Of note is that system dynamics models often generate behavior that is both counterintuitive and at odds with traditional economic theory. Historically, this has caused many system dynamics models to be evaluated critically, especially by some economists. However, today economists from several schools of economic thought are beginning to



use system dynamics, as they have found it useful for incorporating their nontraditional ideas into formal models.

## Introduction

System dynamics is a computer simulation modeling methodology that is used to analyze complex nonlinear dynamic feedback systems for the purposes of generating insight and designing policies that will improve system performance. It was originally created in 1957 by Jay W. Forrester of the Massachusetts Institute of Technology as a method for building computer simulation models of problematic behavior within corporations. The models were used to design and test policies aimed at altering a corporation's structure so that its behavior would improve and become more robust. Today, system dynamics is applied to a large variety of problems in a multitude of academic disciplines, including economics.

System dynamics models are created by identifying and linking the relevant pieces of a system's structure and simulating the behavior generated by that structure. Through an iterative process of structure identification, mapping, and simulation a model emerges that can explain (mimic) a system's problematic behavior and serve as a vehicle for policy design and testing.

From a system dynamics perspective a system's structure consists of stocks, flows, feedback loops, and limiting factors. Stocks can be thought of as bathtubs that accumulate/de-cumulate a system's flows over time. Flows can be thought of as pipe and faucet assemblies that fill or drain the stocks. Mathematically, the process of flows accumulating/de-cumulating in stocks is called integration. The integration process creates all dynamic behavior in the world be it in a physical system, a biological system, or a socioeconomic system. Examples of stocks and flows in economic systems include a stock of inventory and its inflow of production and its outflow of sales, a stock of the book value of a firm's capital and its inflow of investment spending and its outflow of depreciation, and a stock of employed labor and its inflow of hiring and its outflow of labor separations.

Feedback is the transmission and return of information about the amount of information or material that has accumulated in a system's stocks. Information travels from a stock back to its flow(s) either directly or indirectly, and this movement of information causes the system's faucets to open more, close a bit, close all the way, or stay in the same place. Every feedback loop has to contain at least one stock so that a simultaneous equation situation can be avoided and a model's behavior can be revealed recursively. Loops with a single stock are termed minor,

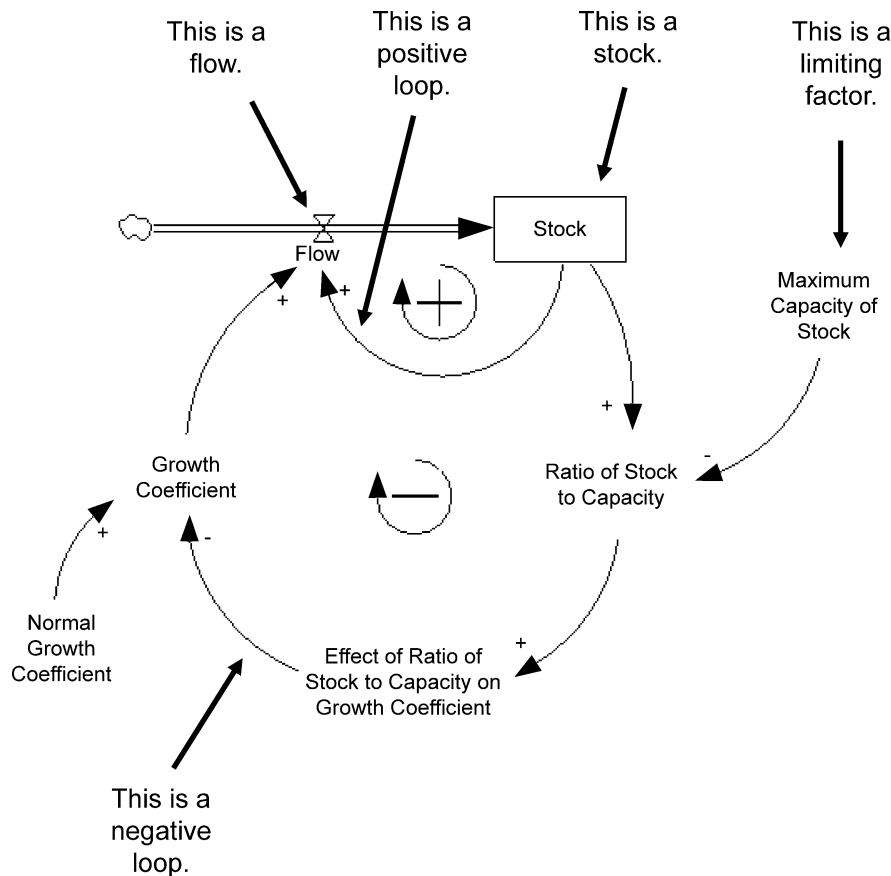
while loops containing more than one stock are termed major.

Two types of feedback loops exist in system dynamics modeling: positive loops and negative loops. Generally speaking, positive loops generate self-reinforcing behavior and are responsible for the growth or decline of a system. Any relationship that can be termed a virtuous or vicious circle is thus a positive feedback loop. Examples of positive loops in economic systems include path dependent processes, increasing returns, speculative bubbles, learning-by-doing, and many of the relationships found in macroeconomic growth theory. Forrester [12], Radzicki and Sterman [46], Moxnes [32], Sterman (Chap. 10 in [55]), Radzicki [44], Ryzhenkov [49], and Weber [58] describe system dynamics models of economic systems that possess dominant positive feedback processes.

Negative feedback loops generate goal-seeking behavior and are responsible for both stabilizing systems and causing them to oscillate. When a negative loop detects a gap between a stock and its goal it initiates corrective action aimed at closing the gap. When this is accomplished without a significant time delay, a system will adjust smoothly to its goal. On the other hand, if there are significant time lags in the corrective actions of a negative loop, it can overshoot or undershoot its goal and cause the system to oscillate. Examples of negative feedback processes in economic systems include equilibrating mechanisms ("auto-pilots") such as simple supply and demand relationships, stock adjustment models for inventory control, any purposeful behavior, and many of the relationships found in macroeconomic business cycle theory. Meadows [27], Mass [26], Low [23], Forrester [12], and Sterman [54] provide examples of system dynamics models that generate cyclical behavior at the macro-economic and micro-economic levels.

From a system dynamics point of view, positive and negative feedback loops fight for control of a system's behavior. The loops that are dominant at any given time determine a system's time path and, if the system is nonlinear, the dominance of the loops can change over time as the system's stocks fill and drain. From this perspective, the dynamic behavior of any economy – that is, the interactions between the trend and the cycle in an economy over time – can be explained as a fight for dominance between the economy's most significant positive and negative feedback loops.

In system dynamics modeling, stocks are usually conceptualized as having limits. That is, stocks are usually seen as being unable to exceed or fall below certain maximum and minimum levels. Indeed, an economic model that can generate, say, either an infinite and/or a negative work-



System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 1  
Simple system dynamics model containing examples of all components of system structure

force would be seen as severely flawed by a system dynamicist. As such, when building a model system dynamicists search for factors that may limit the amount of material or information that the model's stocks can accumulate. Actual socioeconomic systems possess many limiting factors including physical limits (e.g., the number of widgets a machine can produce per unit of time), cognitive limits (e.g., the amount of information an economic agent can remember and act upon), and financial limits (e.g., the maximum balance allowed on a credit card). When limiting factors are included in a system dynamics model, the system's approach to these factors must be described. Generally speaking, this is accomplished with nonlinear relationships. Figure 1 presents a simple system dynamics model that contains examples of all of the components of system structure described above.

### Types of Dynamic Simulation

From a system dynamics point of view, solving a dynamic model – any dynamic model – means determining how

much material or information has accumulated in each of a system's stocks at every point in time. This can be accomplished in one of two ways – analytically or via simulation. Linear dynamic models can be solved either way. Nonlinear models, except for a few special cases, can only be solved via simulation.

Simulated solutions to dynamic systems can be attained from either a continuous (analog) computer or a discrete (digital) computer. Understanding the basic ideas behind the two approaches is necessary for understanding how economic modeling is undertaken with system dynamics.

In the real world, of course, time unfolds continuously. Yet, devising a way to mimic this process on a machine is a bit tricky. On an analog computer, the continuous flow of economic variables in and out of stocks over time is mimicked by the continuous flow of some physical substance such as electricity or water. A wonderful example of the later case is the Phillips Machine, which simulates an orthodox Keynesian economy (essentially the IS-LM model) with flows of colored water moving through pipes and ac-

cumulating in tanks. Barr [2] provides a vivid description of the history and restoration of the Phillips Machine.

On a digital computer, the continuous flow of economic variables in and out of stocks over time is approximated by specifying the initial amount of material or information in a system's stocks, breaking simulated time into small increments, inching simulated time forward by one of these small increments, calculating the amount of material or information that flowed into and out of the system's stocks during this small interval, and then repeating. The solution to the system will always be approximate because the increment of time cannot be made infinitesimally small and thus simulated time cannot be made perfectly continuous. In fact, on a digital computer a trade-off exists between round-off error and integration error. If the increment of time is made too large, the approximate solution can be poor due to integration error. If the increment of time is made too small, the approximate solution can be ruined due to round-off error.

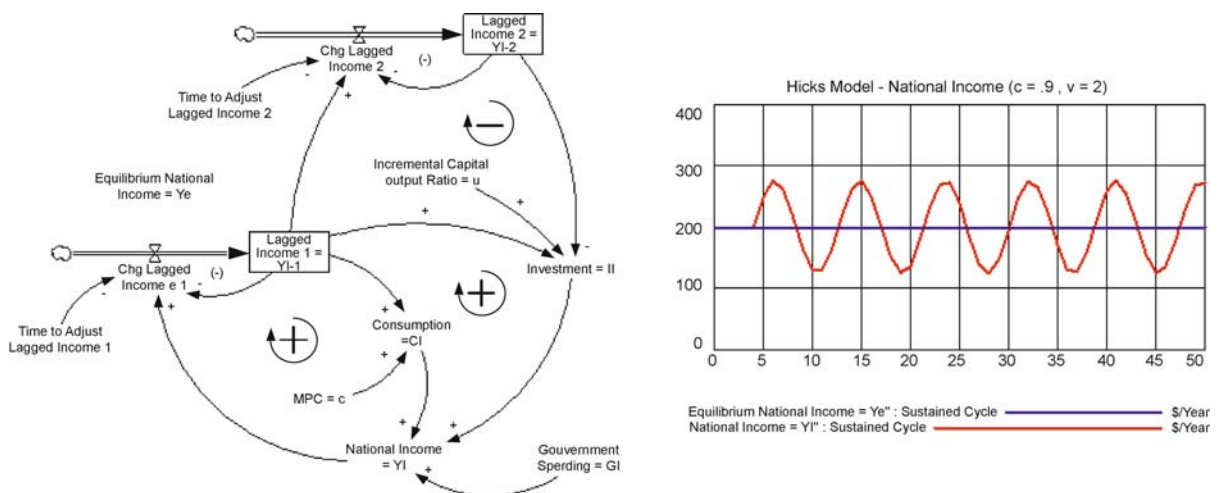
In system dynamics modeling the "true" behavior of the underlying system is conceptualized to unfold over continuous time. As such, mathematically, a system dynamics model is an ordinary differential equation model. To approximate the solution to a continuous time ordinary differential equation model on a digital (discrete) computer, however, difference equations are used. Unlike traditional difference equation modeling in economics, in which the increment of time is chosen to match economic data (typically a quarter or a year), the increment of time in system dynamics modeling is chosen to yield a solution that is accurate enough for the problem at hand, yet avoids

the problems associated with significant round-off and integration error.

The use of difference equations to approximate the underlying differential equations represented by a system dynamics model provides another interesting option when it comes to economic modeling. Since many well known dynamic economic models have been created with difference equations, they can be recast in a system dynamics format by using the difference equations in the system dynamics software literally as difference equations, and not as a tool to approximate the underlying continuous time system. Although doing this deviates from the original ideas embodied in the system dynamics paradigm, it is occasionally done when a modeler feels that analyzing a difference equation model in a system dynamics format will yield some additional insight.

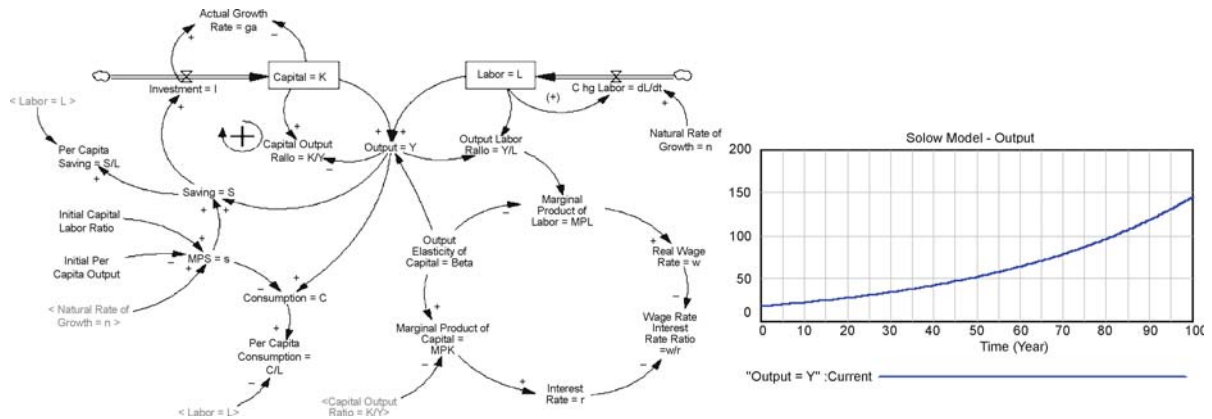
### Translating Existing Economic Models into a System Dynamics Format

There are three principle ways that system dynamics is used for economic modeling. The first involves translating an existing economic model into a system dynamics format, while the second involves creating an economic model from scratch by following the rules and guidelines of the system dynamics paradigm. Forrester [7], Richardson and Pugh [47], Radzicki [42], and Sterman [55] provide extensive details about these rules and guidelines. The former approach is valuable because it enables well-known economic models to be represented in a common format, which makes comparing and contrasting their as-



System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 2

System dynamics representation of John Hicks' multiplier-accelerator difference equation model



**System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 3**  
System dynamics representation of Robert Solow's ordinary differential equation growth model

assumptions, concepts, structures, behaviors, etc., fairly easy. The latter approach is valuable because it usually yields models that are more realistic and that produce results that are “counterintuitive” [11] and thus thought-provoking.

The third way that system dynamics can be used for economic modeling is a “hybrid” approach in which a well known economic model is translated into a system dynamics format, critiqued, and then improved by modifying it so that it more closely adheres to the principles of system dynamics modeling. This approach attempts to blend the advantages of the first two approaches, although it is more closely related to the former.

Generally speaking, existing economic models that can be translated into a system dynamics format can be divided into four categories: written, static (mathematical), difference equation, and ordinary differential equation. Existing economic models that have been created in either a difference equation or an ordinary differential equation format can be translated into system dynamics in a fairly straight-forward manner. For example, Fig. 2 presents Sir John Hicks' [21] Multiplier-Accelerator difference equation model in a system dynamics format and Fig. 3 presents the Robert Solow's [52] ordinary differential equation growth model in a system dynamics format.

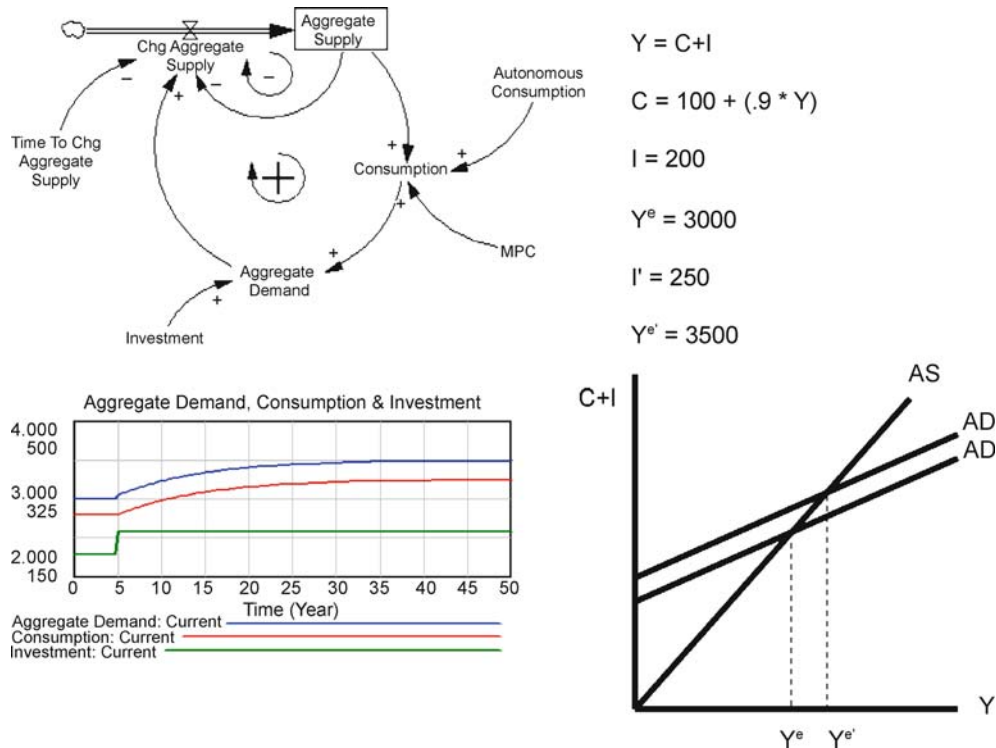
Translating existing static and written economic models and theories into a system dynamics format is a more formidable task. Written models and theories are often dynamic, yet are described without mathematics. Static models and theories are often presented with mathematics, but lack equations that describe the dynamics of any adjustment processes they may undergo. As such, system dynamicists must devise equations that capture the dynamics being described by the written word or that reveal the

adjustment processes that take place when a static system moves from one equilibrium point to another.

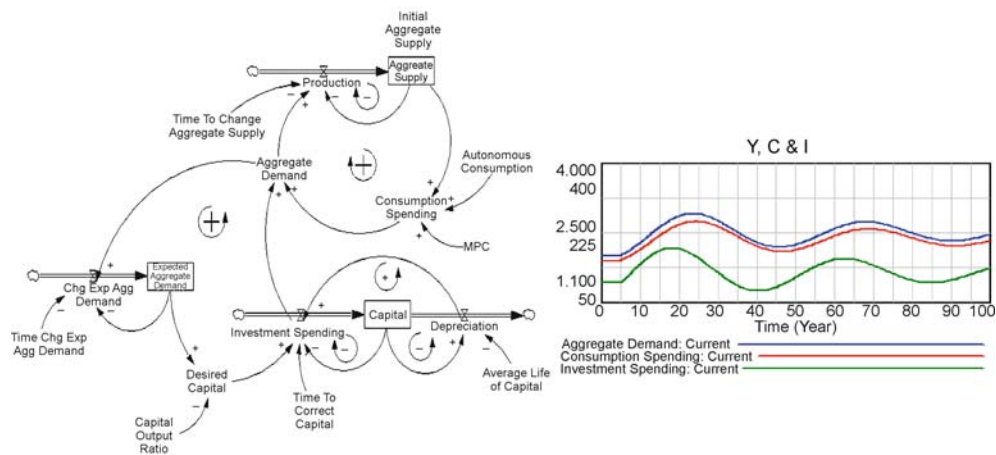
An interesting example of a system dynamics model that was created from a written economic model is Barry Richmond's [48] model of Adam Smith's *Wealth of Nations*. This model was created principally from Robert Heilbroner's [20] written description of Smith's economic system. A classic example of a static model that has been translated into a system dynamics format is a simple two sector Keynesian cross model, as is shown in Fig. 4.

### Improving Existing Economic Models with System Dynamics

The simple two sector Keynesian cross model presented in Fig. 4 is an example of a well known economic model that can be improved after it has been translated into a system dynamics format. More specifically, in this example the flow of investment spending in the model does not accumulate anywhere. This violates good system dynamics modeling practice and can be fixed. Figure 5 presents the improved version of the Keynesian Cross model, which now more closely adheres to the system dynamics paradigm. Other well known examples of classic economics models that have been improved after they have been translated into a system dynamics format and made to conform more closely with good system dynamics modeling practice include the cobweb model [27], Sir John Hicks' multiplier-accelerator model [23], the IS-LM/AD-AS model [13,59], Dale Jorgenson's investment model [51], William Nordhaus' [34] DICE climate change model [4,5], and basic micro economic supply and demand mechanisms [24]. Low's improvement of Hicks' model is particularly interesting because it results



System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 4  
Simple two sector Keynesian cross model in a system dynamics format



System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 5  
Improved simple two sector Keynesian cross model

in a model that closely resembles Bill Phillips' [40] multiplier-accelerator model. Senge and Fiddaman's contributions are also very interesting because they demonstrate how the original economic models are special cases of their more general system dynamics formulations.

### Creating Economic Dynamics Models from Scratch

Although translating well known economic models into a system dynamics format can arguably make them easier to understand and use, system dynamicists believe that the



“proper” way to model an economic system that is experiencing a problem is to do so from scratch while following good system dynamics modeling practice. Unlike orthodox economists who generally follow a deductive, logical positivist approach to modeling, system dynamicists follow an inductive pattern modeling or case study process. More specifically, a system dynamicist approaches an economic problem like a detective who is iteratively piecing together an explanation at a crime scene. All types of data that are deemed relevant to the problem are considered including numerical, written, and mental information. The system dynamicist is guided in the pattern modeling process by the perceived facts of the case, as well as by real typologies (termed “generic structures” in system dynamics) and principles of systems. Real typologies are commonalities that have been found to exist in different pattern models and principles of systems are commonalities that have been found to exist in different real typologies. Paich [36] discusses generic structures at length and Forrester [8] lays out a set of principles of systems.

Examples of a real typologies in economics include Forrester’s [9] *Urban Dynamics* model, which can reproduce the behavior of many different cities when properly parametrized for those cities, and Homer’s [22] model of the diffusion of new medical technologies into the market place, which can explain the behavior of a wide variety of medical technologies when properly parametrized for those technologies. Examples of fundamental principles of systems include the principle of accumulation, which states that the dynamic behavior of any system is due to flows accumulating in stocks, and the notion of stocks and flows being components of feedback loops. The parallels for these principles in economics can be found in modern Post Keynesian economics, in which modelers try to build “stock-flow consistent models,” and in institutional economics, in which the principle of “circular and cumulative causation” is deemed to be a fundamental cause of economic dynamics. Radzicki [41,43,45] lays out the case for the parallels that exist between methodological concepts in system dynamics and methodological concepts in various schools of economic thought.

The economic models that have been historically created from scratch by following the system dynamics paradigm have tended to be fairly large in scale. Forrester’s [12] national economic model is a classic example, as are the macroeconomic models created by Sterman [53], the Millennium Institute [31], Radzicki [45], Wheat [59], and Yamaguchi [60]. Dangerfield [3] has developed a model of Sarawak (E. Malaysia) to analyze and plan for economic transition from a production economy to a knowledge-based one. With the exception

of Radzicki [45], whose model is based on ideas from Post Keynesian and institutional economics, these models, by and large, embody orthodox economic relationships.

### Model Validity

When a system dynamics model of an economic system that is experiencing a problem is built from scratch, the modeling process is typically quite different from that which is undertaken in traditional economics. As such, the question is raised as to whether or not an original system dynamics model is in any sense “valid”.

System dynamicists follow a “pattern modeling” approach [41] and do not believe that models should be judged in a binary fashion as either “valid” or “invalid”. Rather, they argue that confidence in models can be generated along multiple dimensions. More specifically, system dynamicists such as Peterson [38], Forrester and Senge [16] and Barlas [1] have developed a comprehensive series of tests that can be applied to a model’s structure and behavior and they argue that the more tests a model can pass, the more confidence a model builder or user should place in its results. Even more fundamentally, however, Forrester [13] has argued that the real value generated through the use of system dynamics comes, not from any particular model, but from the modeling *process* itself. In other words, it is through the iterative *process* of model conceptualization, creation, simulation, and revision that true learning and insight are generated, and *not* through interaction with the resulting model.

Another issue that lies under the umbrella of model validity involves fitting models to time series data so that parameters can be estimated and confidence in model results can be raised. In orthodox economics, of course, econometric modeling is almost universally employed when doing empirical research. Orthodox economic theory dictates the structure of the econometric model and powerful statistical techniques are used to tease out parameter values from numerical data.

System dynamicists, on the other hand, have traditionally argued that it is not necessary to tightly fit models to time series data for the purposes of parameter estimation and confidence building. This is because:

1. the battery of tests that are used to build confidence in system dynamics models go well beyond basic econometric analysis;
2. the particular (measured) time path that an actual economic system happened to take is merely one of an infinite number of paths that it could have taken and is a result of the particular stream of random shocks that

happened to be historically processed by its structure. As such, it is more important for a model to mimic the basic character of the data, rather than fit it point-by-point [14];

3. utilizing the pattern modeling/case study approach enables the modeler to obtain parameter values via observation below the level of aggregation in the model, rather than via statistical analysis [18];
4. the result of a system dynamics modeling intervention is typically a set of policies that improve system performance and increase system robustness. Such policies are usually feedback-based rules (i. e., changes to institutional structure) that do not require the accurate point prediction of system variables.

Although the arguments against the need to fit models to time series data are well known in system dynamics, many system dynamicists feel that it is still a worthwhile activity because it adds credibility to a modeling study. Moreover, in modern times, advances in software technology have made this process relatively easy and inexpensive. Although several techniques for estimating the parameters of a system dynamics model from numerical data have been devised, perhaps the most interesting is David Peterson's [38,39] Full Information Maximum Likelihood with Optimal Filtering (FIMLOF). Figure 5 presents a run from the Harrod growth model, to which an adaptive expectations structure has been added, after it has been fit via FIMLOF to real GDP and labor supply data for the United States economy for the years 1929–2002. The fit is excellent and the estimated parameter values are consistent with those from more traditional econometric studies. See Radzicki [44] for a detailed description of the model and its parameter estimates.

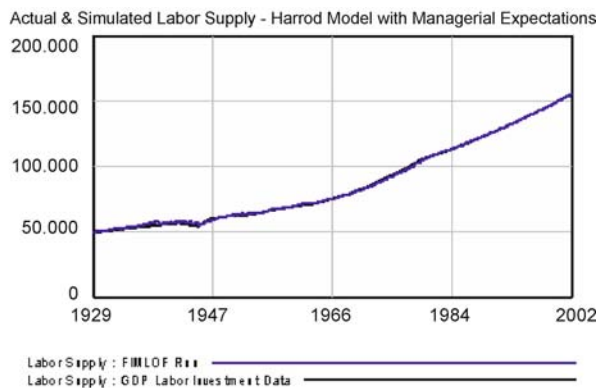
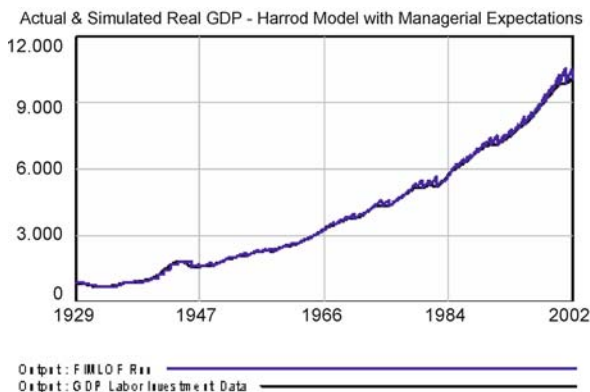
## Controversies

Since system dynamics modeling is undertaken in a way that is significantly different from traditional economic modeling, it should come as no surprise that many economists have been extremely critical of some system dynamics models of economic systems. For example, Forrester's [9] *Urban Dynamics* and [10] *World Dynamics* models have come under severe attack by economists, as has (to a lesser degree) his national economic model. On the other hand, the first paper in the field of system dynamics is Forrester [6], which is essentially a critique of traditional economic modeling.

Greenberger et al. [19] present a nice overview of the controversies surrounding the *Urban Dynamics* and *World Dynamics* models. Forrester and his colleagues' replies to criticisms of the *Urban Dynamics* model are contained in Mass [25] and Schroeder et al. [50].

One of the harshest critics of the *World Dynamics* (WORLD2) model has been Nordhaus [33]. Nordhaus [35] has also very critical of the well known follow-up study to *World Dynamics* known as *The Limits to Growth* [28]. Meadows et al. [29,30] contain updates to the original *Limits to Growth* (WORLD3) model, as well as replies to the world modeling critics.

Forrester [12] presents a nice overview of his national economic model, and the critiques by Stolwijk [57] and Zellner [61] are typical of the attitude of the professional economists toward macroeconomic modeling that is undertaken by following the traditional system dynamics paradigm. The criticism of Forrester's national economic model by the economics profession has probably been less severe, relative to the criticisms of the *Urban Dynamics* and world models, because most of its details are still largely unpublished at the time of this writing.



**System Dynamics and Its Contribution to Economics and Economic Modeling, Figure 6**  
Fit of the Harrod growth model to US macroeconomic data for the years 1929–2002

Another interesting and timely example of the sort of controversy surrounding system dynamics modeling in economics is provided by Sterman and Richardson [56]. In this paper they present a technique for testing whether Hubbert's lifecycle method or the geologic analogy method yields superior estimates of the ultimately recoverable amount of petroleum resources. This study was motivated by a disagreement with a traditionally trained economist over the proper way to conceptualize this issue. Sterman and Richardson devised a clever synthetic data experiment in which a system dynamics model serves as the "real world" with a known ultimately recoverable amount of oil. Hubbert's method and the geologic analogy method are then programmed into the model so they can "watch" the data being generated by the "real world" and provide dynamic estimates of the "known" ultimately recoverable stock of oil. The results showed that Hubbert's method was quite accurate, although it had a tendency to somewhat underestimate the ultimately recoverable amount of oil, while the geologic analogy method tended to overshoot the resource base quite substantially.

### Future Directions

Historically, system dynamicists who have engaged in economic modeling have almost never been trained as professional economists. As such, they have had the advantage of being able to think about economic problems differently from those who have been trained along traditional lines, but have also suffered the cost of being seen as "amateurs" or "boy economists" [41] by members of the economics profession. The good news is that there are currently several schools of economic thought, populated by professional economists, in which system dynamics fits quite harmoniously. These include Post Keynesian economics, institutional economics, ecological economics, and behavioral economics. Historically, the economists in these schools have rejected many of the tenets of traditional economics, including most of its formal modeling methods, yet have failed to embrace alternative modeling techniques because they were all seen as inadequate for representing the concepts they felt were important. However in the modern era, with computers having become ubiquitous and simulation having become in some sense routine, system dynamics is increasingly being accepted as an appropriate tool for use in these schools of economic thought. The future of economics and system dynamics will most probably be defined by the economists who work within these schools of thought, as well as by their students. The diffusion of system dynamics models of economic systems through their translation into user-friendly

interactive "learning environments" that are available over the world wide web will most likely also be of great importance (see [24,59]).

### Bibliography

#### Primary Literature

1. Barlas Y (1989) Multiple Tests for Validation of System Dynamics Type of Simulation Models. *Eur J Operat Res* 42(1):59–87
2. Barr N (1988) The Phillips Machine. *LSE Q Winter* 2(4):305–337
3. Dangerfield BC (2007) System dynamics advances strategic economic transition planning in a developing nation. In: Qudrat-Ullah H, Spector M, Davidsen P (eds) *Complex decision-making: Theory & practice*. Springer, New York, pp 185–209
4. Fiddaman T (1997) Feedback complexity in integrated climate-economy models. Ph D Dissertation, Sloan School of Management, Massachusetts Institute of Technology. Available from <http://www.systemdynamics.org/>
5. Fiddaman T (2002) Exploring policy options with a behavioral climate-economy model. *Syst Dyn Rev* 18(2):243–267
6. Forrester J (1957) Dynamic models of economic systems and industrial organizations. System Dynamics Group Memo D-0. Massachusetts Institute of Technology. Available from <http://www.systemdynamics.org/>
7. Forrester J (1961) *Industrial dynamics*. Pegasus Communications, Inc., Waltham
8. Forrester J (1968) *Principles of systems*. MIT Press, Cambridge
9. Forrester J (1969) *Urban dynamics*. Pegasus Communications, Inc., Waltham
10. Forrester J (1971) *World dynamics*. Pegasus Communications, Inc., Waltham
11. Forrester J (1975) Counterintuitive behavior of social systems. In: Forrester J (ed) *Collected papers of Jay W Forrester*, Pegasus Communications, Inc., Waltham, pp 211–244
12. Forrester J (1980) Information sources for modeling the national economy. *J Am Statist Assoc* 75(371):555–567
13. Forrester J (1985) 'The' model versus a modeling 'process'. *Syst Dyn Rev* 1(1 and 2):133–134
14. Forrester J (2003) Economic theory for the new millennium. In: Eberlein R, Diker V, Langer R, Rowe J (eds) *Proceedings of the Twenty-First Annual Conference of the System Dynamics Society*. Available from <http://www.systemdynamics.org/>
15. Forrester J, Low G, Mass N (1974) The debate on world dynamics: A response to Nordhaus. *Policy Sci* 5:169–190
16. Forrester J, Senge P (1980) Tests for building confidence in system dynamics models. In: Legasto Jr AA, Forrester JW, Lyneis JM (eds) *TIMS Studies in the Management Sciences: System Dynamics*, vol 14. North Holland Publishing Company, Amsterdam, pp 209–228
17. Forrester N (1982) A dynamic synthesis of basic macroeconomic theory: Implications for stabilization policy analysis, Ph D Dissertation, Alfred P Sloan School of Management, Massachusetts Institute of Technology. Available from <http://www.systemdynamics.org/>
18. Graham A (1980) Parameter estimation in system dynamics modeling. In: Randers J (ed) *Elements of the system dynamics method*. Pegasus Communications, Inc., Waltham, pp 143–161
19. Greenberger M, Crenson M, Crissey B (1976) *Models in the policy process: Public decision making in the computer era*, Russell Sage Foundation, New York

20. Heilbroner R (1980) *The worldly philosophers*, 5th edn. Simon & Schuster, New York
21. Hicks J (1950) *A contribution to the theory of the trade cycle*. Oxford University Press, London
22. Homer J (1987) A diffusion model with application to evolving medical technologies. *Technol Forecast Soc Change* 31(3):197–218
23. Low G (1980) The multiplier-accelerator model of business cycles interpreted from a system dynamics perspective. In: Randers J (ed) *Elements of the system dynamics method*. Pegasus Communications, Inc., Waltham, pp 76–94
24. Mashayekhi A, Vakili K, Foroughi H, Hadavandi M (2006) Supply demand world: An interactive learning environment for teaching microeconomics. In: Grosler A, Rouwette A, Langer R, Rowe J, Yanni J (eds) *Proceedings of the of the Twenty-Fourth International Conference of the System Dynamics Society*. Available at <http://www.systemdynamics.org/>
25. Mass N (1974) *Readings in urban dynamics*, vol I. Wright-Allen Press, Cambridge
26. Mass N (1975) *Economic cycles: An analysis of the underlying causes*. MIT Press, Cambridge
27. Meadows D (1970) *Dynamics of commodity production cycles*. Massachusetts MIT Press, Cambridge
28. Meadows D, Meadows D, Randers J, Behrens III W (1972) *The limits to growth: A report for the Club of Rome's project on the predicament of mankind*. Universe Books, New York
29. Meadows D, Meadows D, Randers J (1992) *Beyond the limits: Confronting global collapse, envisioning a sustainable future*. Chelsea Green Publishing, White River Junction
30. Meadows D, Meadows D, Randers J (2002) *Limits to growth: The 30-year update*. Chelsea Green Publishing, White River Junction
31. Millennium Institute (2007) Introduction and Purpose of Threshold 21, <http://www.millennium-institute.org/resources/elibrary/papers/T21Overview.pdf>
32. Moxnes E (1992) Positive feedback economics and the competition between 'hard' and 'soft' energy supplies. *J Sci Ind Res* 51(March):257–265
33. Nordhaus W (1972) World dynamics: Measurement without data. *Econom J* 83(332):1156–1183
34. Nordhaus W (1992) The "DICE" model: Background and structure of a dynamic integrated climate-economy model of the economics of global warming. Cowles Foundation for Research in Economics at Yale University, Discussion Paper No. 1009
35. Nordhaus W (1992) Lethal model 2: The limits to growth revisited. *Brookings Pap Econ Activity* 2:1–59
36. Paich M (1985) Generic structures. *Syst Dyn Rev* 1(1 and 2):126–132
37. Paich M (1994) *Managing the global commons*. MIT Press, Cambridge
38. Peterson D (1975) Hypothesis, estimation, and validation of dynamic social models – energy demand modeling. Ph D Dissertation, Department of Electrical Engineering, Massachusetts Institute of Technology. Available from <http://www.systemdynamics.org/>
39. Peterson D (1980) Statistical tools for system dynamics. In: Randers J (ed) *Elements of the system dynamics method*. Pegasus Communications, Inc., Waltham, pp 226–245
40. Phillips W (1954) Stabilization policy in a closed economy. *Econom J* 64(254):290–323
41. Radzicki M (1990) *Methodologia oeconomiae et systematis dynamis*. *Syst Dyn Rev* 6(2):123–147
42. Radzicki M (1997) Introduction to system dynamics. Free web-based system dynamics tutorial. Available at <http://www.systemdynamics.org/DL-IntroSysDyn/index.html>
43. Radzicki M (2003) Mr. Hamilton, Mr. Forrester and a foundation for evolutionary economics. *J Econom Issues* 37(1):133–173
44. Radzicki M (2004) Expectation formation and parameter estimation in nonergodic systems: the system dynamics approach to post Keynesian-institutional economics. In: Kennedy M, Winch G, Langer R, Rowe J, Yanni J (eds) *Proceedings of the Twenty-Second International Conference of the System Dynamics Society*. Available at <http://www.systemdynamics.org/>
45. Radzicki M (2007) Institutional economics, post keynesian economics, and system dynamics: Three strands of a heterodox economics braid. In: Harvey JT, Garnett Jr. RF (eds) *The future of heterodox economics*. University of Michigan Press, Ann Arbor
46. Radzicki M, Sterman J (1994) Evolutionary economics and system dynamics. In: Englund R (ed) *Evolutionary concepts in contemporary economics*. University of Michigan Press, Ann Arbor, pp 61–89
47. Richardson G, Pugh A (1981) *Introduction to system dynamics modeling with DYNAMO*. Pegasus Communications, Inc., Waltham
48. Richmond B (1985) *Conversing with a classic thinker: An illustration from economics*. Users Guide to STELLA, Chapt 7. High Performance Systems, Inc., Lyme, New Hampshire, pp 75–94
49. Ryzhenkov A (2007) Controlling employment, profitability and proved non-renewable reserves in a theoretical model of the US economy. In: Sterman J, Oliva R, Langer R, Rowe J, Yanni J (eds) *Proceedings of the of the Twenty-Fifth International Conference of the System Dynamics Society*. Available at <http://www.systemdynamics.org/>
50. Schroeder W, Sweeney R, Alfeld L (1975) *Readings in Urban Dynamics*, vol II. Wright-Allen Press, Cambridge
51. Senge P (1980) A system dynamics approach to investment function formulation and testing. *Soc-Econ Plan Sci* 14:269–280
52. Solow R (1956) A contribution to the theory of economic growth. *Q J Econom* 70:65–94
53. Sterman J (1981) *The energy transition and the economy: A system dynamics approach*. Ph D Dissertation, Alfred P Sloan School of Management, Massachusetts Institute of Technology. Available at <http://www.systemdynamics.org/>
54. Sterman J (1985) A behavioral model of the economic long wave. *J Econom Behav Organ* 6:17–53
55. Sterman J (2000) *Business dynamics: Systems thinking and modeling for a complex world*. Irwin-McGraw-Hill, New York
56. Sterman J, Richardson G (1985) An experiment to evaluate methods for estimating fossil fuel resources. *J Forecast* 4(2):197–226
57. Stolwijk J (1980) Comment on 'information sources for modeling the national economy' by Jay W Forrester. *J Am Stat Assoc* 75(371):569–572
58. Weber L (2007) Understanding recent developments in growth theory. In: Sterman J, Oliva R, Langer R, Rowe J, Yanni J (eds) *Proceedings of the of the Twenty-Fifth International Conference of the System Dynamics Society*. Available at <http://www.systemdynamics.org/>



59. Wheat D (2007) The feedback method of teaching macroeconomics: Is it effective? *Syst Dyn Rev* 23(4):391–413
60. Yamaguchi K (2007) Balance of payments and foreign exchange dynamics – S D Macroeconomic Modeling (4). In: Sterman J, Oliva R, Langer R, Rowe J, Yanni J (eds) *Proceedings of the Twenty-Fifth International Conference of the System Dynamics Society*. Available at <http://www.systemdynamics.org/>
61. Zellner A (1980) Comment on ‘information sources for modeling the national economy’ by Jay W Forrester. *J Am Stat Assoc* 75(371):567–569

## Books and Reviews

- Alfeld L, Graham A (1976) *Introduction to urban dynamics*. Wright-Allen Press, Cambridge
- Lyneis J (1980) *Corporate planning and policy design: A system dynamics approach*. PA Consulting, Cambridge
- Meadows D, Meadows D (1973) *Toward global equilibrium: Collected papers*. Wright-Allen Press, Cambridge
- Meadows D, Behrens III W, Meadows D, Naill R, Randers J, Zahn E (1974) *Dynamics of growth in a finite world*. Wright-Allen Press, Cambridge
- Meadows D, Robinson J (1985) *The electronic oracle: Computer models and social decisions*. Wiley, New York
- Richardson G (1999) *Feedback thought in social science and systems theory*. Pegasus Communications, Inc., Waltham
- Sterman J (1988) A skeptic's guide to computer models. In: Grant L (ed) *Foresight and National Decisions*. University Press of America, Lanham, pp 133–169, Revised and Reprinted in: Barney G, Kreutzer W, Garrett M (1991) *Managing a Nation*. Westview Press, Boulder, pp 209–230

## System Dynamics Modeling: Validation for Quality Assurance

MARKUS SCHWANINGER, STEFAN GROESSER  
Institute of Management, University of St. Gallen,  
St. Gallen, Switzerland

### Article Outline

Glossary  
Definition of the Subject  
Introduction  
Epistemological Foundations  
Validation Methods  
Validation Process  
Synopsis and Outlook  
Appendix: Overview of the Tests Described in This Chapter  
Bibliography

### Glossary

**Model/model system** A model is a simplified representation of a real system. Models can be descriptive or pre-

scriptive (normative). Their functions can be to enable explanation, anticipation or design. A distinction used in this contribution is between causal and non-causal models, with System Dynamics models being of the former type. The term *model system* is used to stress the systemic character of a model; this serves to identify it as an organized whole of variables and relationships on the one hand, and to distinguish it from the *real system* which is to be modeled, on the other.

**Model validity** A model's property of adequately reflecting the system modeled. Validity is the primary measure of model quality. It is a matter of degree, not a dichotomized property.

**Model purpose** The goal for which a model is designed or the function it is intended to fulfill. The model purpose is closely linked to the end-model user or model owner. Model purpose is the criterion for the choice of a model's boundary and design.

**Modeling process** The process involving phases such as problem articulation, boundary selection, development of a dynamic hypothesis, model formulation, model testing, policy formulation and policy evaluation [28]. The modeling process is followed by model use and implementation, i. e., the realization of actions designed or facilitated by the use of the model.

**Validation process** Validation is the process by which model validity is enhanced systematically. It consists in gradually building confidence in the usefulness of a model by applying validation tests as outlined in this chapter. In principle, validation pervades all phases of the modeling process, and, in addition, extends into the phases of model use and implementation.

### Definition of the Subject

The present chapter addresses the question of building better models. This is crucial for coping with complexity in general, and in particular for the management of dynamic systems. Both the epistemological and the methodological-technological aspects of model validation for the achievement of high-quality models are discussed. The focus is on formal models, i. e. those formulated in a stringent, logical, and mostly mathematical language.

### Introduction

The etymological root of *valid* is the Latin word *validus*, which denotes attributes such as strong, powerful and firm. A valid model, then, is well-founded and difficult to reject because it accurately represents the perceived real system which it is supposed to reflect. This system can



be either one that already exists or one that is being constructed, or even anticipated, by a modeler or a group of modelers.

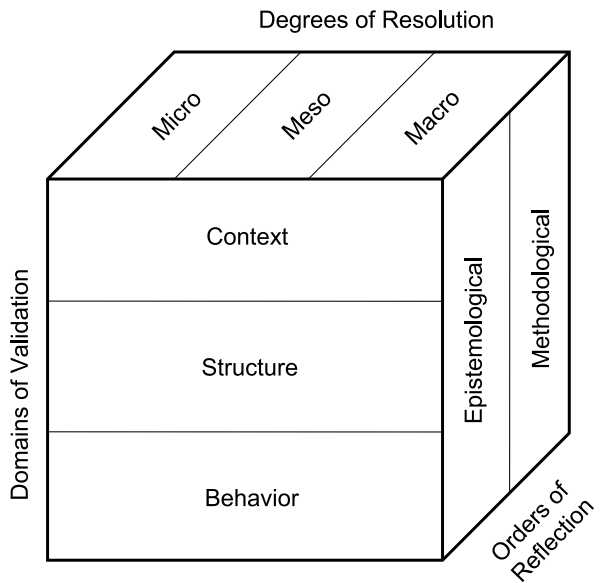
Validation standards in System Dynamics are more rigorous than those of many other methodologies. Let us distinguish between two types of mathematical models, which are fundamentally different: Causal, theory-like models and non-causal, statistical (correlational) models [4]. The former are explanatory, i. e., they embody theory about the functioning of a real system. The latter are descriptive and express observed associations among different elements of a real system. System Dynamics models are causal models.

Non-causal models are tested globally, in that the statistical fit between model and data series from the real system under study is assessed. If the fit is satisfactory, the model is considered to be accurate (“valid”, “true”). In contrast, system dynamicists postulate that models be not only right, but right for the right reasons. As the models are made up of causal interdependencies, accuracy is required for each and every variable and relationship. The following principle applies: if only one component of the model is shown to be wrong, the whole model is rejected even if the overall model output fits the data [4]. This strict standard is conducive to high-quality modeling practice.

A model is an abstract version of a perceived reality. Simulation is a way of experimenting with mathematical models to gain insights and to employ these to improve the real system under study. It is often said that System Dynamics models should portray problems or issues, not systems. This statement must be interpreted in the sense that one should not try to set the boundaries of the model too widely, but rather give the model a focus by concentrating on an object in accordance with the specific purpose of the model. In a narrower definition, even an issue or problem can be conceived of as a “system”, i. e., “a portion of the world sufficiently well defined to be the subject of study” [21]. Validity then consists in a stringent correspondence between model system and real system.

We will treat the issue of model validation as a means of assuring high-quality models. We interject that validity is not the only criterion of model quality, other criteria including parsimony, ease-of-use, practicality, importance, etc. [22].

In the following, the epistemological foundations of model validity are reviewed (Sect. “[Epistemological Foundations](#)”). Then, an overview of the methods for assuring model validity is given (Sect. “[Validation Methods](#)”). Further, the survey includes an overview of the validation process (Sect. “[Validation Process](#)”) and our final conclusions (Sect. “[Synopsis and Outlook](#)”).



**System Dynamics Modeling: Validation for Quality Assurance, Figure 1**

**The Validation Cube – A frame of reference showing three dimensions of the validation topic**

The substance of this article will be made more palpable by means of the following frame of reference. We call it the Validation Cube. The diagram in Fig. 1 shows three dimensions of the validation topic:

- *Orders of Reflection*: We distinguish between an *epistemological* and a *methodological* layer. These define the objects of the next two Sects. “[Epistemological Foundations](#)” and “[Validation Methods](#)”.
- *Domains of Validation*: The three domains, *context*, *structure* and *behavior* refer to the groups of validation methods as described in Sect. “[Validation Methods](#)”.
- *Degrees of Resolution*: We address the different granularities of models. *Micro* refers to the smallest building blocks of models (e. g., variables or small sets of variables), *meso* to modules which constitute a model, and *macro* to the model as a whole.

### Epistemological Foundations

Epistemology is the theory that enquires into the nature and grounds of knowledge: “What can we know and how do we know it?” [13]. These questions are of utmost importance when dealing with models and their validity, because a method of validation is only as good as its epistemological basis.

We can only briefly refer to the antecedents of the epistemological perspective inherent in the idea of model validation as commonly held today in the community of sys-

tem dynamicists. One could go back to Socrates who, in Plato's *Republic* (fourth century BC), addressed the problematic relationship between reality, image and knowledge. One could also refer to John Locke (seventeenth century), the first British empiricist who maintained that ideas could come only from experience, while admitting that our knowledge about external objects is uncertain. We will address the philosophical movements of the nineteenth and twentieth centuries, which are direct sources of the epistemology which is important for model validation. The reader may kindly excuse us for certain massive simplifications that we are obliged to make.

What will be said here about theories applies equally to formal models. In System Dynamics, models either embody theories or they are considered essential components of theories. In addition, processes of modeling and theory-building are of the same nature; a model, like any theory, is built and improved in a dialectic of propositions and refutations [22].

### Positivism and Critique

Positivism is a scientific doctrine founded by Auguste Comte (nineteenth century) which raises the *positive* to the principle of all scientific knowledge. "Positive", in this context, is not meant to be the opposite of negative, but the given, factual, or indubitably existent. The positive is associated with features such as being real, useful, certain, and precise. Positivism confines science to the observable and manipulable, drawing on the mathematical, empirical orientation of the natural sciences as its paragon. The objectivist claim of positivism is that things exist independently of the mind and that truths are detached from human values and beliefs. This stance calls for models that approximate an objective reality.

A younger development in this vein is the school of logical positivism, also logical empiricism (with Schlick, Neurath, Hempel, etc.), which concentrates on the problem of meaning and has developed the verifiability principle: Something is meaningful only if verifiable empirically, i. e., ultimately by observation through the senses. To verify here means to show to be true [13]. For the logical positivists, the method of verification is the essence of theory-building. Tests of theories hinge on their confirmation by facts. In System Dynamics, testing models on real-world data is a core component of validation.

Positivism has been criticized for being reductionist, i. e., for its tendency to reduce concepts to simpler or empirically more accessible ones, and to conceive of learning as an accumulation of particular details. The critique has also asserted that there is no theory-independent identifi-

cation of facts, and therefore different theories cannot be tested by means of the same data [6]. Another objection maintains that social facts are not merely given, but produced by human action, and that they are subject to interpretation [23]. These arguments introduce the principle of relativity, which is of crucial importance for the field of model validation: A model is a subjective construction by an observer.

### Pragmatism – A Challenge to Positivism

Pragmatism, which arose in the second half of the nineteenth century, emphasizes action and the practical consequences of thinking. Its founder, Charles Sanders Peirce, was interested in the effects that the meaning of scientific concepts could have on human experience and action. He defined truth as "the opinion which is fated to be ultimately agreed to by all who investigate" [13], whereby truth is linked to consensual validation. For pragmatists, truth is in what works (Ferdinand Schiller) or satisfies us (John Dewey), and what we find believable and consistent: "The true' ... is only the expedient in the way of our thinking", and "truth is *made* ... in the course of experience." (see p. 581 and p. 583 in [11]).

Pragmatism is often erroneously disdained for supposedly being a crass variety of utilitarianism and embodying a crude instrumentalist rationality. A more accurate view considers the fact that pragmatists are not satisfied with a mere ascertainment of truth; instead they ask: "If an idea or assumption is true, does this make a concrete difference to the life of people? How can this truth be actualized?" In other words, pragmatism does not crudely equate truth and utility. It rather postulates that those truths which are useful to people ought to be put into practice [23].

Pragmatism introduces the criteria of confidence and usefulness, which are more operational as guides to the evaluation of experiments than is the notion of an absolute truth, which is unattainable in the realm of human affairs. At the same time, pragmatism triggers a crucial insight for the context of model-building: The validity of a model depends not only on the absolute quality of that model but also hinges on its suitability with respect to a purpose [7]. In the context of model validation, then, truth is a relative property; more exactly, a *truth* holds for a limited domain only.

### More Challenges to Positivism

We discuss three more challenges to positivism in the twentieth century. First, Thomas Kuhn's theory of scientific revolutions [12]: Kuhn shows, by means of historical cases, that in the sphere of science, generally accepted

ways of looking at the world (“paradigms”) change over time through fundamental shifts. Therefore, the activities of a scientist are largely shaped by the dominant scientific worldview. Second, Willard Van Orman Quine and Wilfrid Sellars argue that knowledge creation and theory-building is a holistic, conversational process, as opposed to the reductionist and confrontational views [4].

Both of these movements contribute to our understanding of how real systems are to be modeled and validated: as organized wholes, and consciously with respect to the values and beliefs underlying a given modeling process. This approach adheres to the spirit of models themselves, by means of which the behavior of whole systems can be simulated and tested on their inherent assumptions.

A third challenge is presented by the interpretive streams of epistemology (for an overview, see [9]). Among them, a main force which expands the possibilities of scientific methodologies is the strand of hermeneutics. Derived from the Greek *hermeneuein* – to interpret or to explain – the term *hermeneutics* stands for a school, mainly associated with Hans-Georg Gadamer, which pursues the ideal of a human science of understanding. The emphasis is on interpretation in an interplay between a subject-matter and the interpreter’s position. This emphasis introduces the subjective into scientific methodology. Hermeneutics denies both that a single “objective true interpretation” can transcend all individual viewpoints, and that humans are forever confined within their own ken [13]. This epistemology offers a necessary complement to a scientific stance, which exclusively hinges on “hard”, quantitative methods in order supposedly to achieve absolute objectivity. The implication of hermeneutics for model validation is that it recognizes the pertinence of subjective judgment. In this connection, interpretive discourses play a crucial role in group model-building and validation. Such discourses lead beyond the subjective, entailing the creation of inter-subjective, shared realities. We will return to this factor in Sect. “Validation Process”.

### Critical Rationalism

Critical rationalism is a philosophical position founded by Karl R. Popper [19,20]. It grew out of positivism but rejected its verificationist stance. Critical rationalism posits that, in the social domain, theories can never be definitely proved, but can only reach greater or lesser levels of truth. Scientific proofs are confined to the realm of the formal sciences, namely logic and mathematics.

As Popper demonstrates, all theories are provisional. As a consequence, the main criterion for the assessment of a theory’s truth status is *falsification* [19]. A theory holds as

long as it is not refuted. Consequently, any theory can be upheld as long as it passes the test of falsification. In other words, the fertile approaches to science are not those of corroboration, but the falsificationist efforts to test if theories can be upheld. In the context of modeling this means that validation must undertake attempts to falsify a model, thereby testing its robustness.

Even Popper’s theory of science is not unchallenged. For example, Kuhn has made the point that its principles are applicable only to normal science, which operates incrementally within a given paradigm, but not to anomalous science, which uncovers unsuspected phenomena in periods of scientific revolution [12]. This observation has an implication for model validation: Alternative and even multiple model designs should be assessed for their ability to account for fundamental change.

### On the Meaning of Validity and Validation

One of the predominant convictions about science is the obsessive idea that proofs are the touchstone of the validity of both theories and models. We follow a different rationale, reverting to the philosophy of science as embodied in critical rationalism.

Popper’s refutationist concept (as opposed to a verificationist concept) of theory-testing implies both an evolutionist perspective and an empiricist stance. The evolutionist perspective is primary because it welcomes the challenges posed to a theory, since these attempts at falsification lead to an evolutionary process: successful falsification efforts result in revisions and improvements of the theory. Correspondingly, empiricism is paramount in the social sciences, because the main source for the refutation of a theory is empirical evidence. However, falsification can also be grounded in logical arguments where empirical evidence cannot be obtained. In this sense, a structuralist approach as used in System Dynamics validation transcends the bounds of logical empiricism.

As a consequence of the evolutionist perspective, there is no such thing as absolute validity. Validity is always imperfect, but it can be improved over time. The empiricist aspect of theory-building implies that theories must be validated by means of empirical data. However, logical assay, estimation and judgment are complementary to this empiricist component (see below).

A validation process is about gradually building confidence in the model under study [2]. This is both analytical and synthetic. It is directed at the model as a whole as much as it is at the components of the model. The touchstone of validity is less whether the model is right or wrong: as Sterman states, “... all models are wrong.” [28].

Some models, however, fulfill the purpose ascribed to them, i. e., they are useful. Models are inherently incomplete; they cannot claim to be true in an absolute sense, but only to be relatively true [4]. In this sense, *validation* is a *goal-oriented* activity and *validity* a *relative* concept.

Finally, the validation process often involves several people because the necessary knowledge is distributed. In these cases, the dialectics of propositions and refutations, as well as the interaction of different subjective viewpoints, and consensus-building, are integral. Validation processes, then, are semiformal, discursive social procedures with a holistic as opposed to a fragmentary orientation [ibidem].

### On Objectivity

If subjective views and judgments are as prominent as alleged above, does objectivity play a role at all? Operational philosophy shows a way out of this dilemma: Rapoport defines objectivity as “invariance with respect to different observers.” [21]. Popper has a similar stance in proposing that general statements must be formulated in a way that they can be criticized and, where applicable, falsified [20]. This concept of objectivity is a challenge to model validation: When defining concepts and functions, one must first of all strive for falsifiable statements. In principle, formal models meet this criterion: each variable and every function or relationship can be challenged. And they must be challenged, so that their robustness can be tested. The duty, then, is in finding the invariances that are inter-subjectively accepted as the best approximations to truth. Frequently this is best achieved in group model-building processes [30]. Finally, truth is something we search for but do not possess [20], i. e., even an accepted model cannot guarantee truth with final certainty.

### Validation Methods

A considerable set of qualitative and quantitative tests has been developed for the enhancement of model validity. The state-of-the-art has been documented in seminal publications [2,4,7,8,14,17,28]. Our purpose here is to present and exemplify the different tests to encourage and help those who strive to develop high-quality System Dynamics models.

In the following, an overview of the types of tests developed for System Dynamics models is given, without any claim to completeness. Most of these tests have been documented extensively in [2,7,8,28]. The descriptions of the tests adhere closely to the specifications of these authors (mainly Forrester and Senge). In addition, we have developed a new category for tests that concentrate on the con-

text in which the model is to be developed. High-quality models can be created only if the relevant context is taken into consideration. To facilitate orientation, we have attached an overview of all described tests in the Appendix.

In this section we describe three groups of tests: those related to model-related context, tests of model structure and tests of model behavior. Many of the tests described in the following can be utilized for explanatory analysis which aims at an understanding of the problematic behavior of the issue under study. Others are suitable for normative ends, in analyzes targeted on improvements of system performance with regard to a specified objective of the reference system. Also known as policy tests, or policy analyses, these “tests of policy implications differ from other tests in their explicit focus on comparing changes in a model and in the corresponding reality. Policy ... tests attempt to verify that response of a real system to a policy change would correspond to the response predicted by the model” [8]. Policy testing can show the risk involved in adopting the model for policy making.

### Tests About the Model-Related Context

These tests deal with aspects related to the situation in which the model is to be developed and embedded. They imply metalevel decisions which have to be taken in the first place, before engaging in model-building. Applied ex-post-facto, i. e., after modeling, they allow for assessing the utility of the modeling endeavor as such.

*Issue Identification Test.* The *raison d'être* of a System Dynamics model is its ability to adequately address an issue and to enhance stakeholders' understanding, an ability which may lead to policy insights and system improvements. The issue identification test examines whether or not the identified issue or problem is indeed meaningful. Has the “right” problem been identified? Does the problem statement address the origins of an issue or only superficial *symptoms*? Whenever complex issues are addressed by a model, different perspectives (e. g. professional, economic, political) must be integrated for accurate problem identification and modeling. This is not a “one-shot-only” test; it must be applied recurrently during the modeling procedure. By reflecting regularly on the correctness of the identified issue, the modeler can increase the likelihood of capturing the origins of suboptimal system behavior.

*Adequacy of Methodology Test.* Simulation models respond to the limitations of humans' mental ability to comprehend complex, dynamic feedback systems [27]. The adequacy of methodology test scrutinizes whether the

System Dynamics methodology is best-suited for dealing with the issue under study. One needs to clearly ascertain if that issue is characterized by dynamic complexity, feedback mechanisms, nonlinear interdependency of structural elements and delays between causes and effects. One needs to ask also if the issue under study could be better addressed by another methodology. For example, in a case where the question is to understand the difference in numerical outcomes between two configurations of a production system, it lets one determine whether discrete event simulation would fulfill this requirement more accurately than System Dynamics.

*System Configuration Test.* This test asks the fundamental question about whether the structural configuration chosen can be accepted. It challenges the assumption that the model represents the actual working of the system under study. The applicability of a different design would be suggested by its ability to capture new conditions, such as different system configurations, phenomena or rules of the game. Even revolutionary changes might be considered. Such an outlook may require a totally new model, or an alternative model designed from a different vantage point. This would at least feasibly approximate the need to take paradigmatic change into account.

*System Improvement Test.* The purpose of modeling is to understand a part of reality and to resolve an issue. The system improvement test can be performed only after the modeling project (an ex-post-facto test), once the insights derived from the model have already been implemented in the real system. This test reestablishes the connection between the abstract mathematical model and the real system. The system improvement test helps to evaluate whether or not model development was successful. In operational terms, any improvements of the real system under study must be compared with explicit objectives. In practice, the test might assess the impact of the modeling process or the model use either on the mental models of decision makers or on changes in organization structures. In principle, assessing the impact of a modeling endeavor is very difficult (one preliminary example is provided by Snabe and Grössler [25]).

## Tests of Model Structure

Tests of model structure refer to the “nuts and bolts” of System Dynamics modeling, i.e., to the formal concepts and interrelationships which represent the real system. Model structure tests aim to increase confidence in the structure of the created theory about the behavior mode of

interest. The model structure can be assessed by means of either direct or indirect inspection. Tests of model structure assess whether the logic of the model is attuned to the corresponding structure in the real world. They do not yet compare the model behavior with time series data from the real system.

**Direct Structure Tests** Direct structure tests assess whether or not the model structure conforms to relevant descriptive knowledge about the real system or class of systems under study. By means of direct comparison, they qualitatively assess any disparities between the original system structure and the model structure.

*Structure Examination Test.* Examination in this case means comparison in the sense just outlined. Qualitative or quantitative information about the real system structure can be obtained either empirically or theoretically. Empirically based tests include reviews of model assumptions about system elements and their interdependencies, e.g., reviews made by highly knowledgeable experts of the real system. Theory-based tests compare the model structure with theoretical knowledge from literature about the type of system being studied. Thereby, a preference for theoretical knowledge specific to the modeled situation over more abstract and general knowledge is usually the case.

To pass the structure examination test, a model must not contradict either the evidence or knowledge about the structure of the real system. This test ensures that the model contains only those structural elements and interconnections that are most likely extant in the real system. In this context, formal inspections of the model's equations, reviews of the syntax for the stock and flow diagram, and walkthroughs along the causal loop diagrams and their embodied causal explanations may be indicated. The experienced reader might recommend the use of statistical tests to identify and validate model structure. As Forrester and Senge [8] indicate, a long-standing discussion exists about the application of inferential statistical tests for structure examination. After a series of experiments, Forrester and Senge conclude “that conventional statistical tests of model structure are not sufficient grounds for rejecting the causal hypotheses in a system dynamics model.” [8]. In the future, however, new statistical approaches might enrich the testing procedures.

*Parameter Examination Test.* A parameter is a quantity that characterizes a system and is held constant in a case under study, but may be varied in different cases (e.g., energy consumption per capita per day). The aim of parameter examination is to evaluate a model's parameters



against evidence or knowledge about the real system. The test can utilize both empirical and theoretical information. Furthermore, the test can be conceptual or numerical. The conceptual parameter examination test is about construct validity; it identifies elements in the real system that correspond to the parameters of the model. Conceptual correspondence means that the parameters match elements of the real system's structure. Numerical parameter examination checks to see if the quantities of the conceptually confirmed parameters are estimated accurately. Techniques for the estimation of parameters are described in [9].

*Direct Extreme Condition Test.* Extreme conditions do not often occur in reality; they are exceptions. The validity of a model's equations under extreme conditions is evaluated by assessing the plausibility of the results generated by the model equations against the knowledge about what would happen under a similar condition in reality. Direct extreme condition testing is a mental process and does not involve computer simulation. Ideally, it is applied to each equation separately. It consists of assigning extreme values to the input variables of each equation. The values of the output variables are then interpreted in terms of what would happen in the real system under these extreme conditions. For example, if a population is zero, then neither births, deaths, nor consumption of resources can occur.

*Boundary Adequacy Structure Test.* Boundary adequacy is given if the model contains the relevant structural relationships that are necessary and sufficient to satisfy a model's purpose. Consequently, the boundary adequacy test inquires whether the chosen level of aggregation is appropriate and if the model includes all relevant aspects of structure. It should ensure that the model contains the concepts that are important for addressing the problem endogenously. For instance, if parameters are likely to change over time, they should be endogenized [8]. The pertinent validation question is: "Should this parameter be endogenized or not?" That question must be decided in view of the model's purpose.

The boundary adequacy test can be applied in three ways: as a structural test, as a behavioral test, and as a policy test. The names are correspondingly: boundary adequacy structure test, boundary adequacy behavior test, and boundary adequacy policy test.

As a test of model structure, the boundary adequacy test involves developing a convincing hypothesis relating the proposed model structure to the particular issue addressed by the model. The boundary adequacy behavior/policy test (explained in Subsect. "Indirect Structure Tests") continues this line of thinking.

*Dimensional Consistency Test.* This test checks the dimensional consistency of measurement units of the expressions on both sides of an equation. The test is performed only at the equation level. When all tests of the individual equations are passed, a large system of dimensionally consistent equations results. This test is passed only if consistency is achieved without the use of parameters that have no meaning in respect to the real world. The dimensional consistency test is a powerful test to establish the internal validity of a model.

**Indirect Structure Tests** Indirect structure tests assess the validity of the model structure indirectly by examining model-generated outcome behaviors. These tests require computer simulation. The comparative activities in these tests are based on logical plausibility considerations which in turn are based on the mental models of the analyst. Comparisons of model generated data and time series about the real system are not yet involved. The tests can be applied to different degrees of model completeness, i. e., to the smallest "atomic" model components, to sub-models, as well as to the entire model.

*Indirect Extreme Condition Test.* For this test, the modeler assigns extreme values to selected model parameters and compares the generated model behavior to the observed or expected behavior of the real system under the same extreme conditions. This test is the logical continuation of the direct extreme condition test, i. e., many of the extreme conditions mentally developed in the previous stage can now be deployed to evaluate the simulated behavioral consequences. This test can be used for the explanatory analysis phase of modeling, but also for the normative phase of policy development. In the first instance, indirect extreme conditions are used to develop a structure that can reproduce the system behavior of interest and guard against developments impossible in reality. In the latter instance, the introduction of policies aims to improve the system's performance. The indirect extreme policy test introduces extreme policies to the model and compares the simulated consequences to what would be the most likely outcome of the real system if the same extreme policies would have been implemented.

*Behavior Sensitivity Test.* Sensitivity analysis assesses changes of model outcome behavior given a systematic variation of input parameters. This test reveals those parameters to which the model behavior is highly sensitive, and asks if the real system would exhibit a similar sensitivity to changes in the corresponding parameters. "The behavior sensitivity test examines whether or not plausi-

ble shifts in model parameters can cause a model to fail behavior tests previously passed. To the extent that such alternative parameter values are not found, confidence in the model is enhanced.” [8]. A model can be numerically sensitive, i. e., the numerical values of variables change significantly, but the behavioral patterns are conserved. It can also exhibit behavioral sensitivity, i. e., the modes of model behavior change remarkably based on systematic parameter variations (Barlas [3] defines several distinct patterns of model behavior).

As the test for indirect extreme conditions, the behavior sensitivity test can also be deployed to assess policy sensitivity. It can reveal the degree of robustness of model behavior and hence indicate to what degree model-based policy recommendations might be influenced by uncertainty in parameter values. If the same policies would be recommended regardless of parameter changes over a plausible range, risk in using the model would be lower than if two plausible sets of parameters lead to distinct policy recommendations.

*Integration Error Test.* Integration error is the deviation between the analytical solution of differential equations and the numerical solution of difference equations. This test ascertains whether the model behavior is sensitive to changes in either the applied integration method or the chosen integration interval (often referred to as simulation time step). Euler’s method is the simplest numerical technique for solving ordinary differential and difference equations. For models that require more precise integration processes, the more elaborated Runge–Kutta integration methods can produce more accurate results, but they require more computational resources.

*Boundary Adequacy Behavior Test/Boundary Adequacy Policy Test.* The logic for testing boundary adequacy has already been developed under the aspect of direct structure testing in the preceding section. The indirect structure version of this test asks whether model behavior would change significantly if the boundary were extended or reduced; i. e., the test involves conceptualizing additional structure or canceling unnecessary structure with regard to the purpose of the study. As one example of expanding the model boundary, this version of the test allows one to detail the treatment of model assumptions considered as unrealistically simple but still important for the model’s purpose. On the other hand, simplifying the model is also a way to reduce the model boundary. The loop-knockout analysis is a useful method to implement this two-sided test. Knockout analysis checks behavior changes induced by the connection and disconnection of a portion of the

model structure, and helps the modeler to evaluate the usefulness of those changes with respect to the model’s purpose.

The other version of this test is the boundary adequacy policy test. It examines whether policy recommendations would change significantly if the boundary were extended (or restricted): That is, what would happen if the boundary assumptions were relaxed (or confined)?

*Loop Dominance Test.* Loop dominance analysis studies the internal mechanisms of a dynamic model and their temporal, relative contribution to the outcome behavior of the model. The relative contribution of a mechanism is a complex quantitative statement that explains the fraction of the analyzed behavior mode caused by the mechanism considered in ► [System Dynamics, Analytical Methods for Structural Dominance Analysis](#) in. The analysis reveals the relative strengths of the feedback loops in the model. The loop dominance test compares these results with the modeler’s or client’s assumption about which are the dominant feedback loops in the real system. Since the results are analytical statements, interpretation and comparison with the real system requires profound knowledge about the system under study.

Loop dominance analysis reveals insights about a model on a different level of analysis than the other validation tests discussed so far: It works not on the level of individual concepts or behaviors of variables but on the level of causal structure, and compares the temporal significance of the different structures to each other. The use of this test for model validation is a novelty. If the relative loop dominances of the model map the relative loop dominances of the real system, confidence in the model is enhanced. If the relative loop dominances of the real system are not known, it is still possible to evaluate whether or not the loop dominance logic in the model is reasonable.

## Tests of Model Behavior

Tests of model behavior are empirical and compare simulation outcomes with data from the real system under study. On that basis, inferences about the adequacy of the model can be made. The empirical data can either be historical or refer to reasonable expectations about possible future developments.

**Behavior Reproduction Tests** The family of behavior reproduction tests examines how well model-generated behavior matches the observed historical behavior of the real system. As a principle, models should be tested against

data not only from periods of stability but also from unstable phases. Policies should not be designed or tested on the premise of normality, but rather should be validated with a view toward robustness and adaptiveness.

**Symptom Generation Test.** This test indicates whether or not a model produces the symptom of difficulty that motivated the construction of the model. To pass the symptom generation test is a prerequisite for considering policy changes, because “unless one can show how internal policies and structures cause the symptoms, one is in a poor position to alter those causes” [8].

Summary statistics, which measure and enable the interpretation of quantitative deviations, provide the means to operationalize the symptom generation test.

One known example is Theil inequality statistics, which measures the mean square-error (MSE) between the model-generated behavior and the historical time series data. It breaks down the deviation into three sources of error: Bias ( $U_m$ ), unequal variation ( $U_s$ ), and unequal co-variation ( $U_c$ ) [26].

An example taken from Schwaninger and Groesser [22] illustrates the interpretation of the error sources.

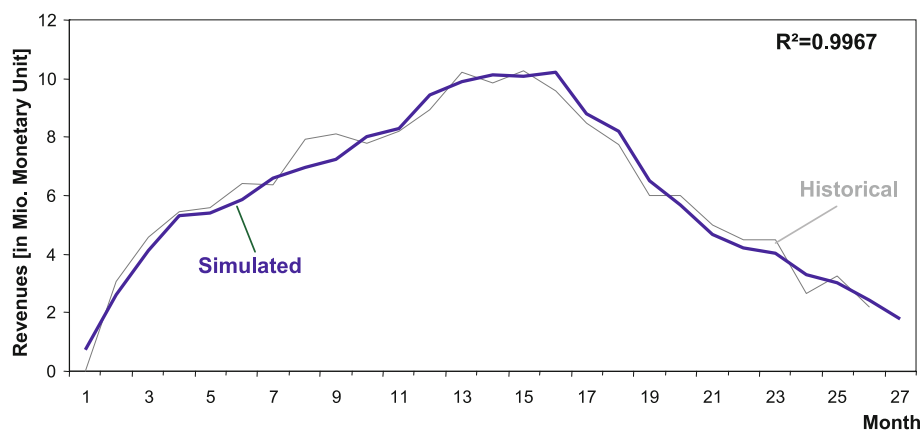
This example from an industrial firm concerns the design of a model that replicates the observed, historical product life-cycle pattern with high accuracy (Fig. 2). “Product Revenue” is the main variable of interest and specifies the symptom (growth phase followed by rapid decay). The mean square-error for revenues is 0.35. The individual components of the inequality statistics are:  $U_m = 0.01$ ,  $U_s = 0.01$ ,  $U_c = 0.98$ . The break down of the statistics shows that the major part of the error is in

the  $U_c$  component, while the other two sources of error are small. This signifies that the point-by-point values of the simulated and historical data do not match, even though the model captures the dominant trend and the average values in the historical data. Such a situation indicates that a major part of the error is probably unsystematic, and therefore the model should not be rejected for failing to match the noise component of the data. The residuals of the historic and simulated time series show no significant trend. This strengthens the assessment that the model comprises a structure that captures the fundamental dynamics of the issue under consideration.

**Frequency Generation and Phase Relationship Tests.** These tests focus on the frequencies of time series and phase relationships between variables. An example is the pattern of investment cycles in an industry. These tests are superior to point-by-point comparisons between model-generated and observed behavior (cf. [7]).

Frequency refers to periodicities of fluctuation in a time series. Phase relationship is the relationship between the time series of at least two variables. In principle, three phase relations are possible: Preceding, simultaneous, and successive. The frequency generation test evaluates whether or not the periodicity of a variable is in accordance with the real system. The phase relationship test assesses the phase shifts of at least two variables by comparing their trajectories.

If the phase shift between the selected simulation variables contradicts the phase shift between the same variables as observed or expected in the real system, a structural flaw in the model might be diagnosed. The test can



System Dynamics Modeling: Validation for Quality Assurance, Figure 2

An example comparison of historical and simulated time series for product revenues. The explained variance is close to 100% ( $R^2 = 0.9967$ )

uncover failures in the model, but offers only little guidance as to where the erroneous part of the model might be. The autocorrelation function test is one way to operationalize the frequency generation test [1]. The function test consists in comparing the autocorrelation functions of the observed and the model-generated behavior outputs, and can detect if significant errors between them exist.

**Modified Behavior Test.** Modified behavior can arise from a modified model structure or changes in parameter values. This test concerns changes in the model structure. It can be performed if data about the behavior of a structurally modified version of the real system are available. “The model passes this test if it can generate similar modified behavior, when simulated with structural modifications that reflect the structure of the “modified” real system” [2]. The applicability of this test is rather limited since it requires specific data about the modified real system which must be similar in kind to the original real system. Only under this condition can additional insights into the suitability of the original model structure be obtained. If the modified real system deviates strongly from the original real system, the test does not result in any additional insights, because no stringent conclusions about the validity of the original system can be derived from a model that is dissimilar in its structure.

**Multiple Modes Test.** A mode is a pattern of observed behavior. The multiple mode test considers whether a model is able to generate more than one mode of observed behavior, for instance, if a model about the production sector of an economy generates distinct patterns of fluctuations for the short-term (production, employment, inventories, and prices) and for the long term (investment, capital stock) [15]. “A model able to generate two distinct periodicities of fluctuation observed in a real system provides the possibility for studying possible interaction of the modes and how policies differentially affect each mode” [8].

**Behavior Characteristic Test.** Characteristics of a behavior are features of historical data that are clearly distinguishable, e.g., the peculiar shape of an oscillating time series, sharp peaks, long troughs, or such unusual events as an oil crisis. Since System Dynamics modeling is not about point prediction, the behavior characteristic test evaluates whether or not the model can generate the circumstances and behavior leading to the event. The creation of the exact time of the behavior is not part of the test.

**Behavior Anticipation Tests** System Dynamics models do not strive to forecast future states of system variables.

Nevertheless, given that the fundamental system structure is not subject to rapid and fundamental change, dynamic models might provide insights about the possible range of future behaviors. Hence, behavior anticipation tests are similar to behavior reproduction tests but possess a higher level of uncertainty.

**Pattern Anticipation Test.** This test examines whether a model generates patterns of future behavior which are assumed to be qualitatively correct. The limits of anticipation reside in the fact that the structure of the system may change over time. The pattern anticipation test entails evaluation of periods, phase relationships, shape, or other characteristics of behavior anticipated by the model. One possibility for implementing this test is to split the historical time series into two data sets and introduce an artificial present time at the end of the first data series. The first set is then used for model development and calibration. The second data series is employed to perform the behavior anticipation test, i.e., to evaluate whether the model is able to anticipate possible future behavior.

This test can also be used for policy considerations, in which case it is called “Changed Behavior Anticipation Test”. It determines whether the model correctly anticipates how the behavior of the real system will change if a governing policy is altered.

**Event Anticipation Test.** In respect to System Dynamics, the anticipation of events does not imply knowing the exact time at which the events occur; it rather means understanding the dynamic nature of events and being able to identify the antecedents leading to them. For instance, the event anticipation test is passed if a model has the ability to anticipate a steep peak in food prices based on the development of the conditioning factors.

**Behavior Anomaly Test** In constructing and analyzing a System Dynamics model, one strives to make it behave like the real system under study. However, the analyst may detect anomalous features of the model’s behavior which conflict with the behavior of the real system. Once the behavioral anomaly is traced to components of the model structure responsible for the anomaly, one often finds flaws in model assumptions. The test for recognizing behavioral anomalies is sporadically applied throughout the modeling process.

**Family Member Test** A System Dynamics model often represents a family of social systems. Whenever possible, a model should be a general representation of the class of that system to which the particular case belongs. One

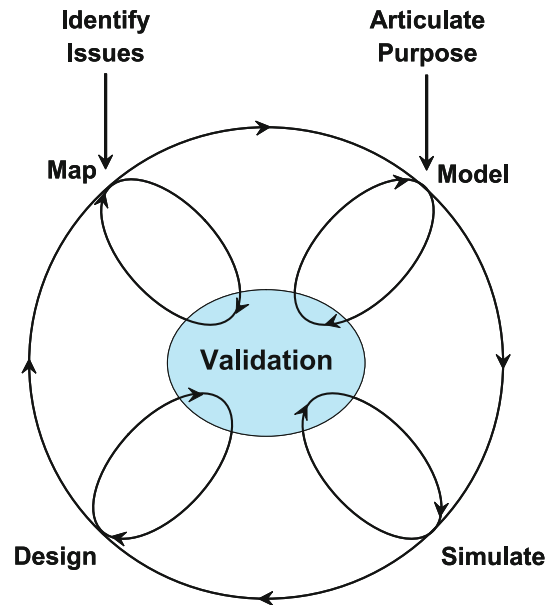
should ask if the model can generate the behavior in other instances of the same class. “The family-member test permits a repeat of the other tests of the model in the context of different special cases that fall within the general theory covered by the model. The general theory is embodied in the structure of the model. The special cases are embodied in the parameters. To perform this test, one uses the particular member of the general family for picking parameter values. Then one examines the newly parametrized model in terms of the various model tests to see if the model has withstood transplantation to the special case” [8]. The model should be calibrated so as to be applicable to the widest range of related systems. For the family member test, only the parameter values of the model are subject to alterations; changes in the model structure are part of the modified behavior test, as discussed in the preceding section.

**Surprise Behavior Test** A surprising model behavior is a behavior that is not expected by the analysts. When such an unexpected behavior appears, the model analysts must first understand the causes of the unexpected behavior within the model. They then compare the behavior and its causes with those of the real system. In many cases, the surprising behavior turns out to be due to a formulation flaw in the model. However, if this procedure leads to the identification of behavior previously unrecognized in the real system, the confidence in the model’s usefulness is strongly enhanced. Such a situation may signify a model-based identification of a counter-intuitive behavior in a social system.

**Turing Test** The Turing test is a qualitative test which uses the intuitive knowledge of system experts to evaluate model behavior. Experts are presented with a shuffled collection of real and simulated output behavior patterns. They are asked if they can distinguish between these two types of patterns. If they are unable to discern which pattern belongs to the real system and which to the simulation output, the Turing test is passed. Similar to the phase relationship test, the Turing test is powerful in its ability to indicate structural flaws, but offers only little guidance for locating them in the model.

### Validation Process

The validation process pervades all phases of model-building and reaches even beyond, into the phases of model implementation and use. The diagram in Fig. 3 visualizes the function of validation in the process of model-building.



System Dynamics Modeling: Validation for Quality Assurance, Figure 3

Validation in the context of the System Dynamics modeling procedure

For the purposes of this contribution, validation is placed at the center of the scheme. From there it is dispersed through all steps of the modeling process, Map (high-level model creation), Model (build the formal model), Simulate (explore scenarios, etc.) and Design (articulation of policies). We have limited the differentiation of these steps in order to highlight the structure of the process – a recursive structure drawn as a nested loop line. After the initial identification of issues and the articulation of model purpose, the simplified diagram denotes the four phases, of mapping to modeling to simulation and design. The small loops symbolize micro-processes in which, for example, a model is submitted to validation, e. g., a direct structure test, which may lead to its modification (two small arrows). The larger loops illustrate more comprehensive processes. For example, an indirect structure test of the model is carried out, in which the behavior is tested by means of simulation. Or a policy test by simulation leads to implications for design (large loop), and the design is validated in detail thereafter (small loop).

Now, we should note that the process scheme reminds us of a further aspect which is quite fundamental. If the results of the model’s operation, e. g., a “prediction”, diverge from the results of a test, then either the model is wrong or the test is inadequate (see p. 168 in [24]). This meta-perspective lets us keep an eye on the adequacy of the tests:



is the logic of the test flawless? Are the data sources in order? (see adequacy of methodology test in Sect. “[Validation Methods](#)”).

Model-building is a process of knowledge-creation, and model validation is an integral part of it. As the model is validated using the methods described in the former chapter, insights emerge, and a better understanding of the system under study keeps growing. But model-building is also a construction of a reality in the minds of observers [31,32] concerned with an issue. In this procedure, validation is supposed to be a “guarantor” for the realism of the model, a control function for preventing gross aberrations in individual and collective perceptions. Validation should encompass precautions against cognitive limitations and modeler blindness. The set of tests presented above is a system of heuristic devices for enhancing such provisions. A question not yet answered is how these tests should be ordered along the timeline. We have fleshed out three structural principles, which are illustrated in Fig. 4:

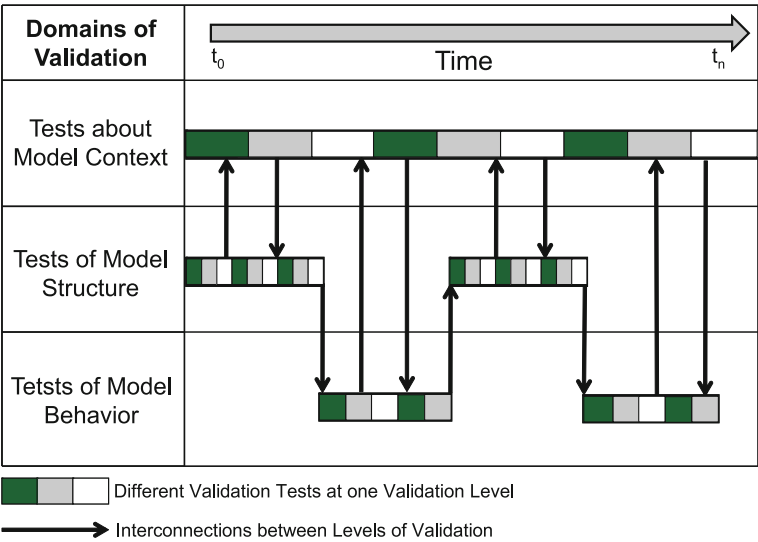
1. *Validation is a parallel process:* Validation in all three domains – context, structure and behavior – is carried out in a synchronized fashion, as shown in Fig. 4. Context validation is continuous, while the other two components show alternations.
2. *Parts of the validation process have a sequential structure:* This refers to the alternations between the components of structure and behavior validation. In principle, they occur alternately, with structural validation taking the lead and behavior validation following. After that,

one might revert to structural validation again and so forth.

3. *Validation processes are polyrhythmic:* The length and accentuation of validation activities vary among the three levels. This fact is symbolized by the frequency of the vertical lines in the blocks of the chronogram.

A further important factor affecting the validation process is the degree of resolution: micro, meso or macro (as visualized in Fig. 1). The focus of validation is primarily on micro-objects, the smallest building blocks of a model, for example, a stock or a subsystem containing a stock with its flows. One could call them metaphorically *atoms* or *molecules*. Each building-block should be validated individually, before it is integrated into the overall model structure. The reason is that at this atomic level disfunctionalities or errors of thinking are discovered immediately, while at higher levels of resolution the identification of structural flaws is more difficult and cumbersome. The same holds for the relation between modules (meso) and the whole model (macro). Before adding a module, it should be validated in itself. This way, errors at the level of the whole system can be minimized and, it is very important to add, counterintuitive behavior of the model can be understood with more ease.

Until now we have examined what occurs in a validation process and how the process is structured. Finally, we raise the issue of who the actors are and why. In this context, we will concentrate on group processes in model validation.



System Dynamics Modeling: Validation for Quality Assurance, Figure 4  
The interplay of validation activities

Different observers associate diverse contents with a system, and they might even conceive the system distinctly, as far as its boundaries, goals and structures are concerned. They might also succumb to erroneous inferences and therefore adhere to defective propositions. Consequently, error-correcting devices are needed. A powerful mechanism for this purpose is the practice of model-building and validation in groups. We have already referred to that concept in respect to several of the methods discussed in Sect. “[Validation Methods](#)”, and now we will briefly expand on it.

Group Model-building (GMB) is a methodology to facilitate team learning with the help of System Dynamics [30]. The methodology consists of a set of methods and instruments as well as heuristic principles. These are meant to facilitate the elicitation of knowledge, the negotiation of meanings, the creation of a shared understanding of a problem in a team, as well as the joint construction and validation of models. The process of GMB is essentially a dialog in which different interpretations of the real system under study are exposed, transformed, aligned and translated into the concepts and relationships which make up the model system. This is mainly a matter of structural validation, of qualitative mapping and the elaboration of the formal model.

Given its transdisciplinary approach, GMB enables an integration of different perspectives into one shared image of the system-in-focus. GMB is an important provision for attaining higher model quality: it can broaden the available knowledge base, inhibit errors and show itself to be a cohesive force in the quest for consensual model validation. The opportunity for validation inheres in the broad knowledge base normally available in a modeling group. Much of this knowledge can be leveraged for validation purposes. Most validation tests are carried out in coordination with model-building activities. Often the tests become a task to be accomplished between workshops. However, the members of the model-building group can, in principle, be made available for knowledge input into and monitoring of validation activities.

A functioning GMB process requires a number of necessary elements [18]: commitment of key players (e.g., attendance of workshops), impartial facilitation, on-the-spot modeling at conversational pace, with continuous display of the developing model as well as an interactive and iterative group process.

Let us not forget that there are many situations in which one single person is in charge of building and validating a model. In these cases the modeler must constantly challenge his or her own position. Normally, it is preferred that one should also call for external judgment in reviews,

walkthroughs and the like. The same holds for knowledge supply. One-person modelers can find a lot of material in the media, libraries, the internet, etc., but it is also usually beneficial to find experienced persons from whom to elicit relevant knowledge, or even persons who join the modeling and validation venture.

## Synopsis and Outlook

Models should be relevant for coping with the complexity of the real world. At the same time, the methods by which they are constructed must be rigorous; otherwise the quality of the model suffers. Rigor and relevance are not entirely dichotomous, but given resource constraints they are in competition to a certain extent. Lack of rigor in building a model is often worse than limitations to the model's relevance. One may say, *cum grano salis*: incomplete validation entails complete irrelevance. Modelers must find a way to ensure both rigor and relevance, as both are necessary conditions for achieving the model purpose. Neither alone is sufficient, but one may assume that, taken together, rigor and relevance are sufficient conditions. The relative importance of these two dimensions of model building may vary over time as a function of the model quality achieved. At the beginning, relevance might be more important, while at high levels of model accomplishment rigor might become prevalent.

Investing in high model quality is indeed both worthwhile and imperative. It is impressive to register the fact that model validation has achieved higher levels of rigor not only in the academic field but also in the world of affairs: According to Coyle and Exelby, the need for orientating decisions about “real-world” affairs has also fueled strong efforts among commercial modelers and consultants for ensuring model validity [5].

We have discussed two essential aspects of model validation, the epistemological foundations and methodological procedures for ensuring model validity. The main conclusion we have reached on epistemology is that crude positivism has been superseded by newer philosophical orientations that provide guidance for an adequate concept of validation in System Dynamics. Validation has been defined as a rich and well-defined process by which the confidence in a model is gradually enhanced. Validity, then, is always a matter of degree, never an absolute property.

*Well-defined* here is not meant in the sense of a rigid algorithm, but as the rigorous application of a battery of validation methods which we have described in some detail. We have included a number of new validation tests by which modelers' understanding of the relevant con-

text can be scrutinized. These additional tests are rightly supposed to prevent wrong methodological choices. They should also trigger innovative approaches to the issues under study and foster the ability to think in terms of contingencies. Finally, they should liberate modelers from tunnel vision and open avenues to creativity. The imperative here is to cultivate a “sense of the possible” (Robert Musil’s *Möglichkeitssinn*) and a skepticism against the supposedly impossible (see also [29]).

Simulation based on formal dynamic models is likely to become ever more important for both private and public organizations. It will continue to support managers at all levels in decision-making and policy design. The more that models are relied upon, the greater the importance of their high quality. Therefore, model validation is one of the big issues lying ahead in System Dynamics modeling.

## Appendix: Overview of the Tests Described in This Chapter

1. **Tests of the Model-Related Context**
  - 1.1 Issue Identification Test
  - 1.2 Adequacy of Methodology Test
  - 1.3 System Configuration Test
  - 1.4 System Improvement Test
2. **Tests of Model Structure**
  - 2.1 Direct Structure Tests
    - 2.1.1 Structure Examination Test
    - 2.1.2 Parameter Examination Test
    - 2.1.3 Direct Extreme Condition Test
    - 2.1.4 Boundary Adequacy Structure Test
    - 2.1.5 Dimensional Consistency Test
  - 2.2 Indirect Structure Tests
    - 2.2.1 Indirect Extreme Condition Test
    - 2.2.2 Behavior Sensitivity Test
    - 2.2.3 Integration Error Test
    - 2.2.4 Boundary Adequacy Behavior Test/Boundary Adequacy Policy Test
    - 2.2.5 Loop Dominance Test
3. **Tests of Model Behavior**
  - 3.1 Behavior Reproduction Tests
    - 3.1.1 Symptom Generation Test
    - 3.1.2 Frequency Generation and Phase Relationship Test
    - 3.1.3 Modified Behavior Test
    - 3.1.4 Multiple Modes Test
    - 3.1.5 Behavior Characteristic Test
  - 3.2 Behavior Anticipation Tests
    - 3.2.1 Pattern Anticipation Test
    - 3.2.2 Event Anticipation Test
  - 3.3 Behavior Anomaly Test

- 3.4 Family Member Test
- 3.5 Surprise Behavior Test
- 3.6 Turing Test

## Bibliography

### Primary Literature

1. Barlas Y (1990) An autocorrelation function test for output validation. *Simulation* 55(1):7–16
2. Barlas Y (1996) Formal aspects of model validity and validation in system dynamics. *Syst Dyn Rev* 12(3):183–210
3. Barlas Y (2006) Model validity and testing in System Dynamics: Two specific tools. Paper presented at the 24th International Conference of the System Dynamics Society, Nijmegen
4. Barlas Y, Carpenter S (1990) Philosophical roots of model validity – two paradigms. *Syst Dyn Rev* 6(2):148–166
5. Coyle G, Exelby D (2000) The validation of commercial system dynamics models. *Syst Dyn Rev* 16(1):27–41
6. Feyerabend P (1993) *Against method*, 3rd edn. Verso, London
7. Forrester JW (1961) *Industrial dynamics*. MIT Press, Cambridge
8. Forrester JW, Senge PM (1980) Test for building confidence in System Dynamics models. In: Legasto AA Jr, Forrester JW, Lyneis JM (eds) *System Dynamics*. North-Holland Publishing Company, Amsterdam, pp 209–228
9. Graham AK (1980) Parameter estimation in system dynamics modeling. *TIMS Studies in the Management Sciences* 14:125–142
10. Heracleous L (2006) *Discourse, interpretation, organization*. Cambridge University Press, Cambridge
11. James W (1987) *Writings 1902–1910*. Library of America, New York
12. Kuhn T (1996) *The structure of scientific revolutions*, 3rd edn. University of Chicago Press, Chicago
13. Lacey AR (1996) *A dictionary of philosophy*, 3rd revised edn. Barnes and Noble, New York
14. Lane DC (1995) The folding star: A comparative reframing and extension of validity concepts in System Dynamics. In: Simada T, Saeed K (eds) *Proceedings of 1995 international System Dynamics conference*, 30 July–4 Aug, vol I. System Dynamics Society, Lincoln, pp 111–130
15. Mass NJ (1975) *Economic cycles: An analysis of underlying causes*. Productivity Press, Cambridge
16. Mattheij RMM, Rienstra SW, Boonkkamp JH MtT (2005) *Partial differential equations: Modeling, analysis, computation*. Society for Industrial and Applied Mathematics (SIAM), Eindhoven
17. Petersen DW, Eberlein RL (1994) Understanding models with vensim. In: Morecroft JDW, Sterman JD (eds) *Modeling for learning organizations*. Productivity Press, Portland, pp 339–358
18. Phillips LD (2007) Decision conferencing. In: Edwards W, Miles RF, von Winterfeldt D (eds) *Advances in decision analysis. From foundations to applications*. Cambridge University Press, Cambridge, pp 375–399
19. Popper KR (1959) *The logic of scientific discovery*. Basic Books, New York (latest edition: 2002, Routledge, London)
20. Popper KR (1972) *Objective knowledge: An evolutionary approach*. Clarendon Press, Oxford
21. Rapoport A (1954) *Operational philosophy. Integrating knowledge and action*. Harper, New York

22. Schwaninger M, Groesser SN (2008) Model-based theory-building with system dynamics. *Syst Res Behav Sci* 25:1–19
23. Seiffert H, Radnitzky G (1994) *Handlexikon der Wissenschaftstheorie*, 2nd edn. DTV Wissenschaft, Munich
24. Smith VL (2008) *Rationality in economics: Constructivist and ecological forms*. Cambridge University Press, Cambridge
25. Snabe B, Grössler A (2006) System dynamics modelling for strategy implementation – case study and issues. *Syst Res Behav Sci* 23(4):467–481
26. Sterman JD (1984) Appropriate summary statistics for evaluating the historical fit of system dynamics models. *Dynamica* 10(2):51–66
27. Sterman JD (1989) Misperceptions of Feedback in Dynamic Decision Making. *Organ Behav Human Decis Process* 43(3):301–335
28. Sterman JD (2000) *Business dynamics. Systems thinking and modeling for a complex world*. Irwin/McGraw-Hill, Boston
29. Taleb NN (2007) *The black swan. The impact of the highly improbable*. Random House, New York
30. Vennix JAM (1996) *Group model building: Facilitating team learning using System Dynamics*. Wiley, Chichester
31. von Foerster H (1984) *Observing systems*, 2nd edn. Intersystems Publications, Seaside
32. von Glasersfeld E (1991) Abschied von der Objektivität. In: Watzlawick P, Krieg P (eds) *Das Auge des Betrachters*. Piper, Munich, pp 17–30

### Books and Reviews

- Finlay PN (1997) Validity of decision support systems: Towards a validation methodology. *Syst Res Behav Sci* 14(3):169–182
- Forrester JW (1961) *Industrial dynamics*. MIT Press, Cambridge
- Law AM (2007) *Simulation modeling and analysis*, 4th edn. McGraw-Hill, New York
- Legasto AA Jr, Forrester JW, Lyneis JM (eds) (1980) *System Dynamics*. North-Holland, Amsterdam
- Morecroft J (2007) *Strategic modelling and business dynamics: A feedback systems approach*. Wiley, Chichester
- Sargent RG (2004) Validation and verification of simulation models. In: Ingalls RG, Rossetti MD, Smith JS, Peters BA (eds) *Proceedings of the 2004 winter simulation conference*. ACM-Association for Computing Machinery, Washington DC, pp 17–28
- Sterman JD (2001) *Business dynamics. Systems thinking and modeling for a complex world*. Irwin/McGraw-Hill, Boston
- Warren K (2008) *Strategic management dynamics*. Wiley, Chichester

## System Dynamics Models of Environment, Energy and Climate Change

ANDREW FORD

School of Earth and Environmental Sciences,  
Washington State University, Pullman, Washington, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[The Model of Mono Lake](#)

[The Model of the Salmon in the Tucannon River](#)

[Models of Climate Change](#)

[System Dynamics Models of the Carbon Cycle](#)

[Lessons from the Regulated Power Industry in the 1970s](#)

[Simulating the Power Industry Response to a Carbon Market](#)

[Conditions for Effective Interdisciplinary Modeling](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**CO<sub>2</sub>** Carbon dioxide is the predominant greenhouse gas. Anthropogenic CO<sub>2</sub> emissions are created largely by the combustion of fossil fuels.

**CGCM** Coupled general circulation model, a climate model which combines the atmospheric and oceanic systems.

**GCM** General circulation model, a term commonly used to describe climate models maintained at large research centers.

**GHG** GHG is a greenhouse gas such as CO<sub>2</sub> and methane. These gases contribute to global warming by capturing some of the outgoing infrared radiation before it leaves the atmosphere.

**GT** Gigaton, a common measure of carbon storage in the global carbon cycle. A GT is a billion metric tons.

**IPCC** The Intergovernmental Panel on Climate Change was formed in 1988 by the World Meteorological Organization and the United Nations Environmental Program. It reports research on climate change. Their assessments are closely watched because of the requirement for unanimous approval by all participating delegates.

### Definition of the Subject

System dynamics is a methodology for studying and managing complex systems which change over time. The method uses computer modeling to focus our attention on the information feedback loops that give rise to the dynamic behavior. Computer simulation is particularly useful when it helps us understand the impact of time delays and nonlinearities in the system. A variety of modeling methods can aid the manager of complex systems. Coyle (p. 2 in [3]) puts the system dynamics approach in perspective when he describes it as that “branch of control theory which deals with socio-economic systems, and that

branch of management science which deals with problems of controllability.” The emphasis on controllability can be traced to the early work of Jay Forrester [9] and his background in control engineering [10]. Coyle highlighted controllability again in the following, highly pragmatic definition:

*System dynamics is a method of analyzing problems in which time is an important factor, and which involve the study of how a system can be defended against, or made to benefit from, the shocks which fall upon it from the out-side world.*

The emphasis on controllability is important as it directs our attention to understanding and managing the system, not to the goal of forecasting the future state of the system. Making point predictions is the objective of some modeling methods, but system dynamics models are used to improve our understanding of the general patterns of dynamic behavior. System dynamics has been widely used in business, public policy and energy and environmental policy making. This article describes applications to energy and environmental systems.

## Introduction

System dynamics has been used extensively in the study of environmental and energy systems. This article describes some of these applications, paying particular attention to the problem of global climate change. The applications were selected to illustrate the power of the method in promoting an interdisciplinary understanding of complex problems.

The applications to environmental and energy systems are similar to applications to other systems described in this encyclopedia. They usually begin with the recognition of a dynamic pattern that represents a problem. System dynamics is based on the premise that we can improve our understanding of the dynamic behavior by the construction and testing of computer simulation models. The models are especially helpful when they illuminate the key feedbacks that give rise to the problematic behavior.

System dynamics is explained in the core article in this volume, in the early texts by Forrester [9], Coyle [3] and Richardson [18] and in more recent texts on strategy by Warren [22] and by Morecroft [17]. The most comprehensive explanation is provided in the text on business dynamics by Sterman [19]. Applications to environmental systems are explained in the text by Ford [7]. The most widely read application to the environment is undoubtedly *The Limits to Growth* [16]. Collections of environmental

applications appear in special issues of the *System Dynamics Review* [11,20].

The models are normally implemented with visual software such as Stella (<http://www.iseesystems.com/>), Vensim (<http://www.vensim.com/>) or Powersim (<http://www.powersim.com/>). These programs use stock and flow icons to help one see where the accumulations of the system take place. They also help one to see the information feedback in the simulated system. The programs use numerical methods to show the dynamic behavior of the simulated system. The examples selected for this article make use of the Stella and Vensim software.

This article begins with textbook examples of environmental resources in the western US. The management of water levels at Mono Lake in Northern California is the first example. It shows a hydrological model to simulate the decline in lake levels due to water exported out of the basin. The second example involves the declining salmon population in the Tucannon River in Eastern Washington. These examples demonstrate the clarity of the approach, and they illustrate the potential for interdisciplinary modeling.

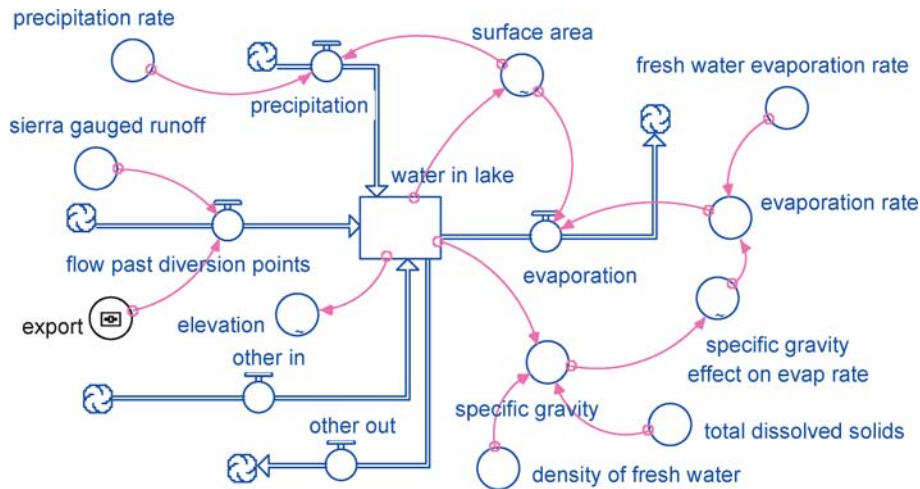
The article then turns to the topic of climate change and global warming. The focus is on the global carbon cycle and the growing concentration of carbon dioxide (CO<sub>2</sub>) in the atmosphere. A wide variety of models have been used to improve our understanding of the climate system and the importance of anthropogenic CO<sub>2</sub> emissions. Examples of system dynamics models are presented to show how they can improve our understanding and provide a platform for interdisciplinary analysis.

System dynamics has also been widely applied in the study of energy problems, especially problems in the electric power industry. The final section describes two applications to electric power. The first involved the financial problems of regulated electric utilities in the US during the 1970s. It demonstrates the usefulness of the method in promoting an interdisciplinary understanding of the utilities' financial problems. The second study dealt with the CO<sub>2</sub> emissions in the large electricity system in the Western USA and Canada. It demonstrated how the power industry could lead the way in reducing CO<sub>2</sub> emissions in the decades following the implementation of a market in carbon allowances.

## The Model of Mono Lake

Mono Lake is an ancient inland sea on the east side of the Sierra Nevada Mountains in California. Microscopic algae thrive in its saline waters, and the algae support huge populations of brine flies and brine shrimp which can,





System Dynamics Models of Environment, Energy and Climate Change, Figure 1  
Stella diagram of the model of Mono Lake

under the right conditions, provide a virtually limitless food supply for migratory and nesting birds. Starting in 1941, stream flows toward Mono Lake were diverted into the aqueduct for export to Los Angeles. The large export deprived the lake of the historical flows, and the volume shrunk over the next four decades. By 1980, the lake's volume was cut approximately in half, and its salinity nearly doubled. Higher salinity levels posed risks to the ecosystem, and environmental scientists feared for the future of the lake ecosystem. Various groups filed suit in the 1970s to limit exports, and the California Supreme Court ruled in 1983 that public trust doctrine mandated a reconsideration of the management of the waters of the Mono Basin. That reconsideration led to a long-term plan to limit exports until the lake's elevation would return to safer levels.

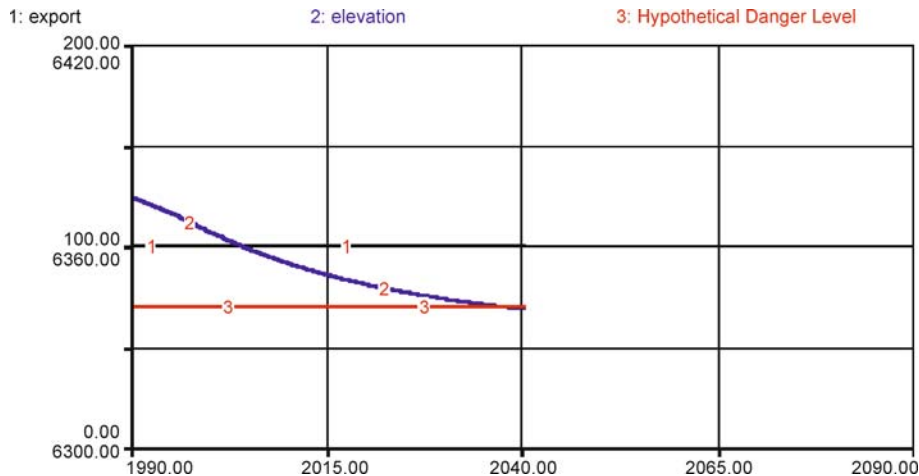
Figure 1 shows a system dynamics model to simulate water flows and storage in the Mono Basin. The goal was to understand the pattern of decline over four decades and to study the responsiveness of the lake to a change in export policy. The model is implemented with the Stella software, and Fig. 1 shows how the model appears when using the software. A single stock variable is used to represent the storage in the basin. The main flow into the lake is the flow from gauged streams that bring runoff from the Sierra to the lake. The aqueduct system diverts a portion of this flow south to Los Angeles, and the flow allowed past the diversion points is the main flow into the lake. The main outflow is the evaporation. It depends on the surface area of the lake and the evaporation rate. The surface area depends in a nonlinear way on the volume of water in the lake. Figure 1 shows that this model follows the standard, system dynamics practice of using familiar names to con-

vey the meaning of the variables in the model. (These particular names match the terms used by water managers and hydrological models of the basin.)

Figure 2 shows the simulated decline in the lake if exports were allowed to continue at high levels for 50 years. The lake would decline from 6374 to around 6342 feet above sea level, a value which is designated as a hypothetical danger level for this simulation. The long, gradual decline is a match of projections by the other hydrological models used in the management plan for the basin. The lake will continue to fall until the area has been reduced sufficiently to create an evaporation which will lead to a balance of the flows in and out of the basin.

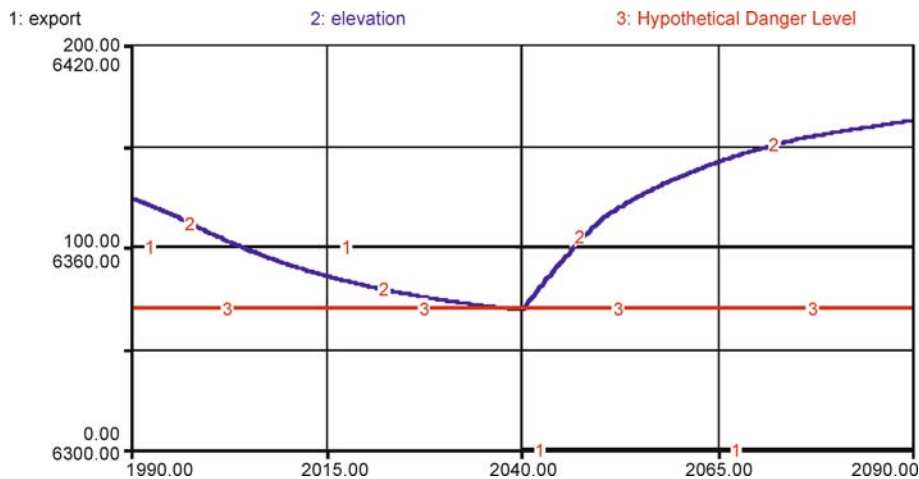
Figure 3 shows the simulated responsiveness of the lake to a change in export. The export is cut to zero midway through the simulation, and the elevation increases rapidly in the ensuing decade. The simulation reveals an immediate and rapid response, indicating that there is little downward momentum associated with the hydrology of the basin. This responsiveness is highly relevant to the management plan. When the lake falls to a dangerous level, the export could be reduced, and the lake would climb to higher elevations within a few years after the change in policy. This rapid response supports the "wait and see" argument by those who advocated waiting for full signs of a dangerous salinity before changing export policy.<sup>1</sup> But there is far more than hydrology at work in this system. The waters of Mono Lake support a com-

<sup>1</sup>"Wait and see" may be supported by an analysis of the hydrology of the basin, but it does not necessarily make sense when considering the long delays in the political and managerial process to change water export.



System Dynamics Models of Environment, Energy and Climate Change, Figure 2

Simulated decline in Mono Lake elevation if historical export were allowed to continue until the year 2040



System Dynamics Models of Environment, Energy and Climate Change, Figure 3

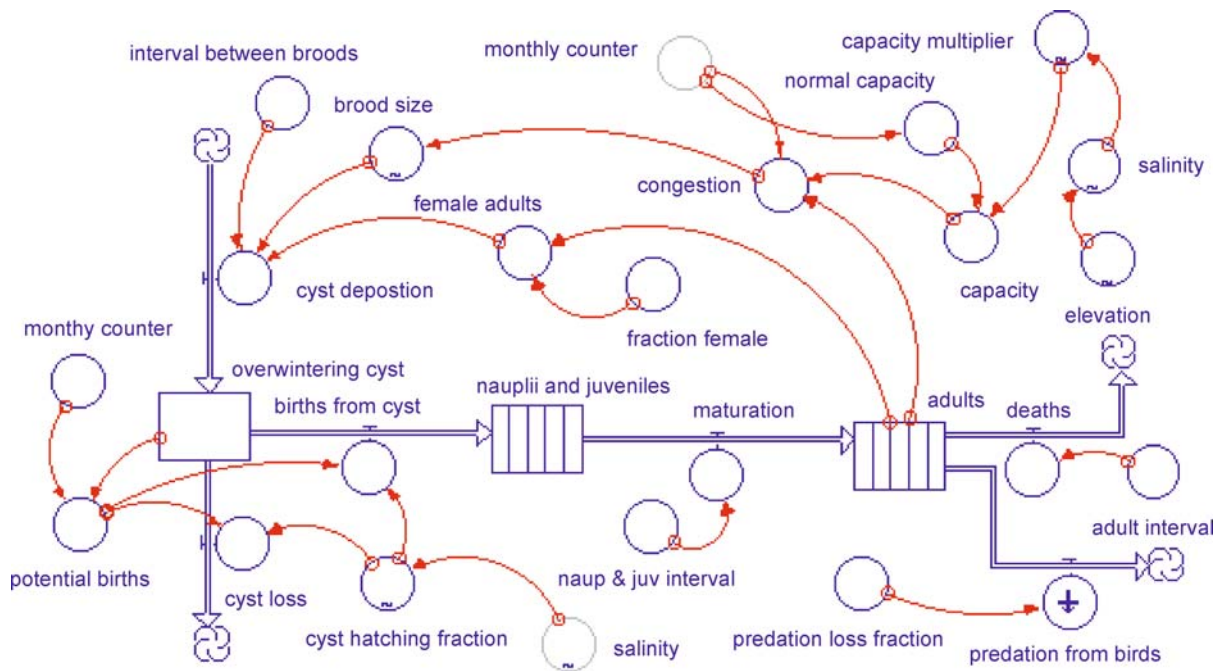
Simulated recovery of Mono Lake elevation if export is set to zero for the second half of the simulation

plex ecosystem which may or may not recover as quickly as the lake elevation. To explore the larger system requires an interdisciplinary model, one that looks at both hydrology and population biology.

Figure 4 shows a model of the population of brine shrimp that live in Mono Lake. The life cycle begins when the adult females deposit cysts in the summer. A stock is assigned to the over wintering cysts. The nauplii and juvenile phases are combined into a second stock, and the maturation leads to a new population of adults in the following summer. The model operates in months and is simulated over a long time interval to show the population response to long-term changes in elevation and in salinity. The model shows the population's response to changes in lake elevation, so one can learn about the delays in the pop-

ulation's response to the changes in lake elevation. Since the shrimp life cycle is 12 months, one would expect the population to rebound rapidly after the increase in elevation and the reduction in salinity. The model confirms that the shrimp population would increase rapidly in the years following the elimination of water export from the basin.

The Mono Lake models are textbook models [7]. They demonstrate the clarity that the system dynamics approach brings to the modeling of environmental systems. The stock and flow icons help one see the structure of the system, and the long variable names help one appreciate the individual relationships. The simulation results help one understand the downward momentum in the system. In this particular case, there is no significant down-



**System Dynamics Models of Environment, Energy and Climate Change, Figure 4**  
**Stella model of the brine shrimp population of Mono Lake**

ward momentum associated with either the hydrological dynamics or the population dynamics.

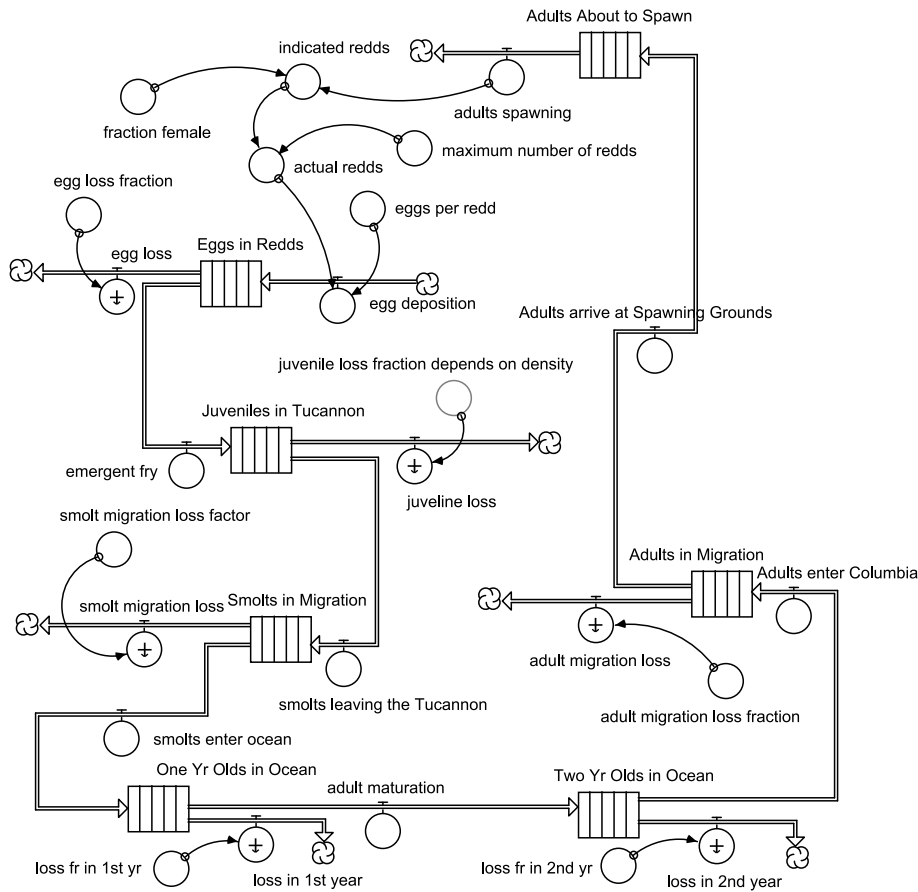
The model in Fig. 1 allows for a system dynamics portrayal of the type of calculations commonly performed by hydrologists. Compared to the previous methods in hydrology, system dynamics adds clarity and ease of experimentation. The population model in Fig. 4 is a system dynamics version of the type of modeling commonly performed by population biologists. System dynamics adds clarity and ease of experimentation in this discipline as well.

The main theme of this article is that system dynamics offers the opportunity for interdisciplinary modeling and exploration. The Mono Lake case illustrates this opportunity with the combination of the hydrological and biological models that allows one to simulate management policies that control export based on the size of the brine shrimp population. The new model is no longer strictly hydrology nor strictly population biology; it is an interdisciplinary combination of both. And by using stock and flow symbols that are easily recognized by experts from many fields of study, the system dynamics enables quick transfer of knowledge. The ability to combine perspectives from different disciplines is one of the most useful aspects of the system dynamics approach to environmental and energy systems. This point is illustrated further with each of the remaining examples in the article.

### The Model of the Salmon in the Tucannon River

The next example involves the decline in salmon populations in the Snake and Columbia River system of the Pacific Northwest. By the end of the 1990s, the salmon had disappeared from 40% of their historical breeding ranges despite a public and private investment of more than \$1 billion. The annual salmon and steelhead runs had dwindled to less than a quarter of the runs from one hundred years ago. Figure 5 shows a system dynamics model one of the salmon runs, the population of Spring Chinook that spawn in the Tucannon River. The river rises in the Blue Mountains of Oregon and flows 50 miles toward the Snake River in Eastern Washington. It is estimated that the river originally supported runs of 20 thousand adults. But the number of returning adults has declined substantially due to many changes in the past sixty years. These changes include agricultural development in the Tucannon watershed, hydro-electric development on the Snake and Columbia, and harvesting in the ocean.

Each of the stocks in Fig. 5 correspond to a different phase in the salmon life cycle (see Table 1), with a total life-cycle of 48 months. The parameters represent predevelopment conditions, the conditions prior to agricultural development in the Tucannon watershed and hydro-electric development on the Snake and Columbia. Each of these parameters is fixed regardless of the size of the salmon



System Dynamics Models of Environment, Energy and Climate Change, Figure 5  
Stella diagram of the model of the salmon life cycle

populations. One of the most important variables is the “juvenile loss fraction depends on density.” It can be as low as 50% when there are only a few emergent fry each spring. With higher densities, however, juvenile survival becomes more difficult due to crowding in the cool and safe portions of the river.

Figure 6 shows the model results over a 480 month period with the population parameters in Table 1. The simulation begins with a small number to see if the population will grow to the 20 thousand adults that were thought to have returned to the river in earlier times. The time graph shows a rapid rise to around 20 thousand adults within the first 120 months of the simulation. The remainder of the simulation tests the population response to variability in environmental conditions, as represented by random variations in the smolt migration loss fraction. (This loss tends to be high in years with low runoff and low in years with high runoff.) Figure 6 confirms that the model simulates the major swings in returning adults due to environmental

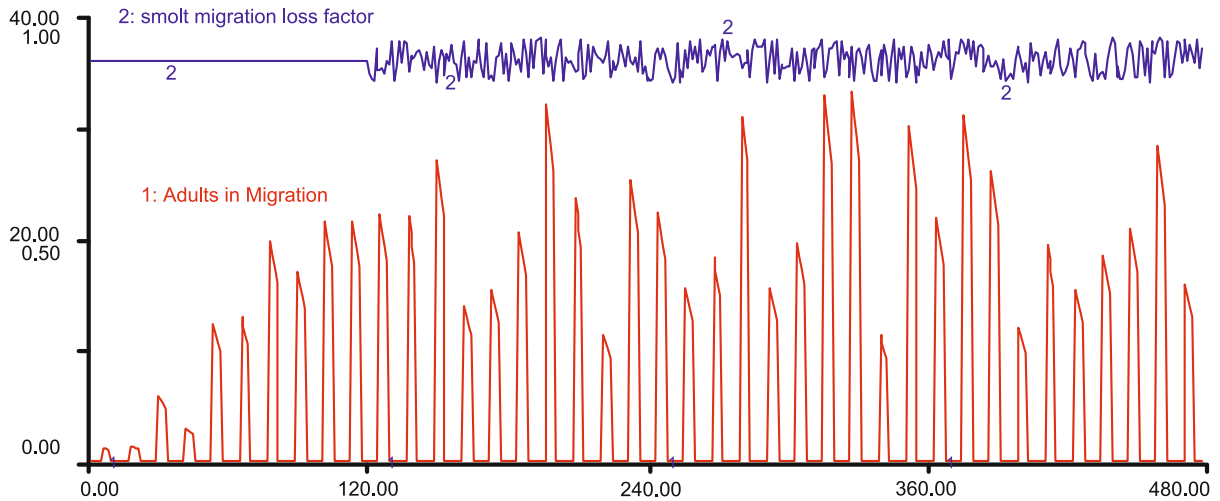
variability. The runs can vary from a low of ten thousand to a high of thirty thousand.

System dynamics models are especially useful when they help us to understand the key feedbacks in the system. Positive feedback loops are essential to our under-

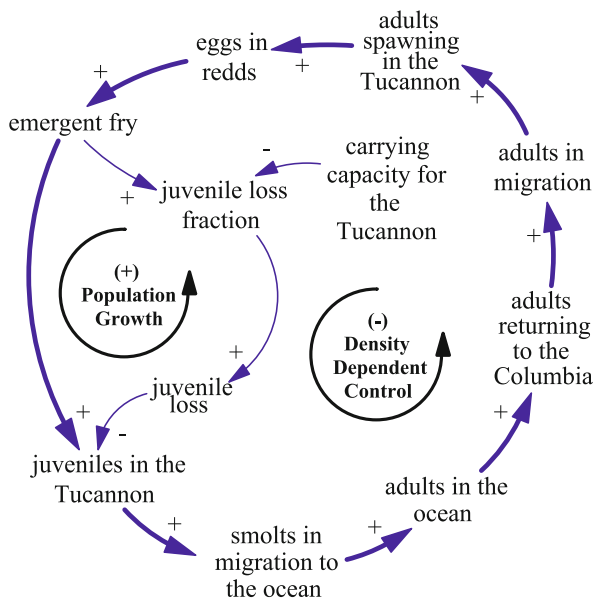
System Dynamics Models of Environment, Energy and Climate Change, Table 1

Inputs to simulate the salmon population under pre-development conditions

| Months in each phase  |    | Population parameters         |       |
|-----------------------|----|-------------------------------|-------|
| Adults ready to spawn | 1  | fraction female               | 50%   |
| eggs in redds         | 6  | eggs per redd                 | 3,900 |
| juveniles in Tucannon | 12 | egg loss fraction             | 50%   |
| smolts in migration   | 1  | smolt migration loss factor   | 90%   |
| one yr olds in ocean  | 12 | loss fr for first yr          | 35%   |
| two yr olds in ocean  | 12 | loss fr for second yr         | 10%   |
| adults in migration   | 4  | adult migration loss fraction | 25%   |



System Dynamics Models of Environment, Energy and Climate Change, Figure 6  
Test of the salmon model with random variations in the smolt migration losses



System Dynamics Models of Environment, Energy and Climate Change, Figure 7  
Key feedback loops in the salmon model

standing of rapid, exponential growth; negative feedbacks are essential to our understanding of the controllability of the system. Causal loop diagrams are often used to depict the feedback loops at work in the simulated system. Figure 7 shows an example by emphasizing the most important feedback loops in the salmon model.

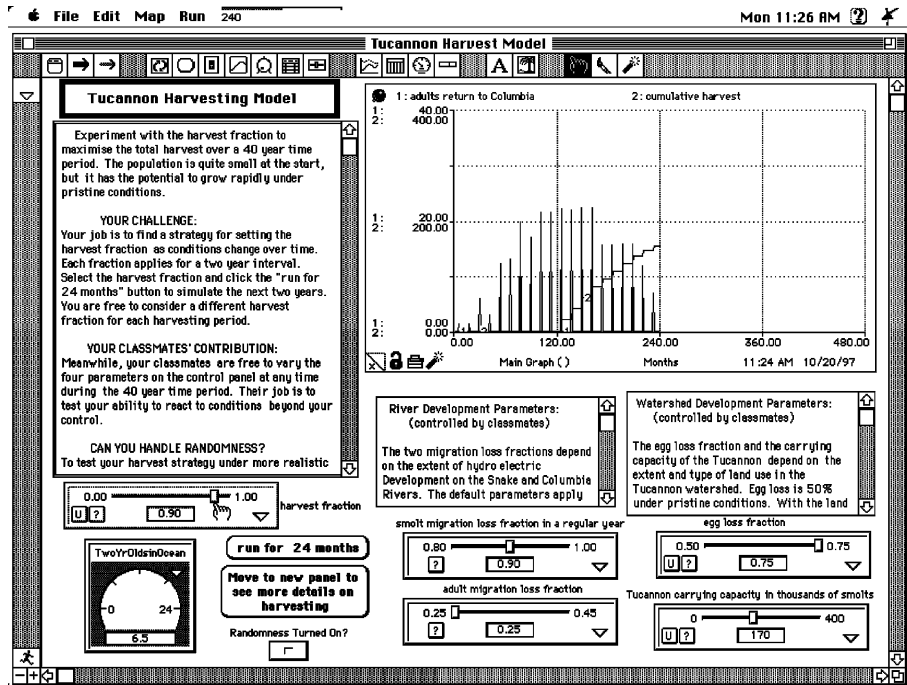
Most readers will immediately recognize the importance of the outer loop which is highlighted by bold arrows in the diagram. Starting near the top, imagine that there

are more spawning adults and more eggs in redds. We would then expect to see more emergent fry, more juveniles, more smolts in migration, more salmon in the ocean, more adults entering the Columbia, and a subsequent increase in the number of spawning adults. This is the positive feedback loop that gives the salmon population the opportunity to grow rapidly under favorable conditions.

An equally important feedback works its way around the inner loop in the diagram. If we begin at the top with more spawners, we would expect to see more eggs, more fry and a greater juvenile loss fraction as the fry compete for space in the river. With a higher loss fraction, we expect to see fewer juveniles survive to be smolts, fewer smolts in migration, and fewer adults in the ocean. This means we would see fewer returning adults and less egg deposition. This “density dependent feedback” becomes increasingly strong with larger populations, and it turns out to be crucial to the eventual size of the population. Simulating density dependent feedback is also essential to our understanding of the recovery potential of the salmon population. Suppose, for example, that the salmon experience high losses during the adult migration. This will mean that fewer adults reach the spawning grounds. There will be less egg deposition and fewer emergent fry in the following spring. The new cohort of juveniles will then experience more favorable conditions, and a larger fraction will survive the juvenile stage and migrate to the ocean. The density dependent feedback is crucial to the population’s ability to withstand shocks from external conditions.<sup>2</sup>

<sup>2</sup>The shocks could take the form of changes in ocean mortalities, changes in harvesting and changes in the migration mortalities. These





System Dynamics Models of Environment, Energy and Climate Change, Figure 8  
Salmon harvesting model to encourage student experimentation

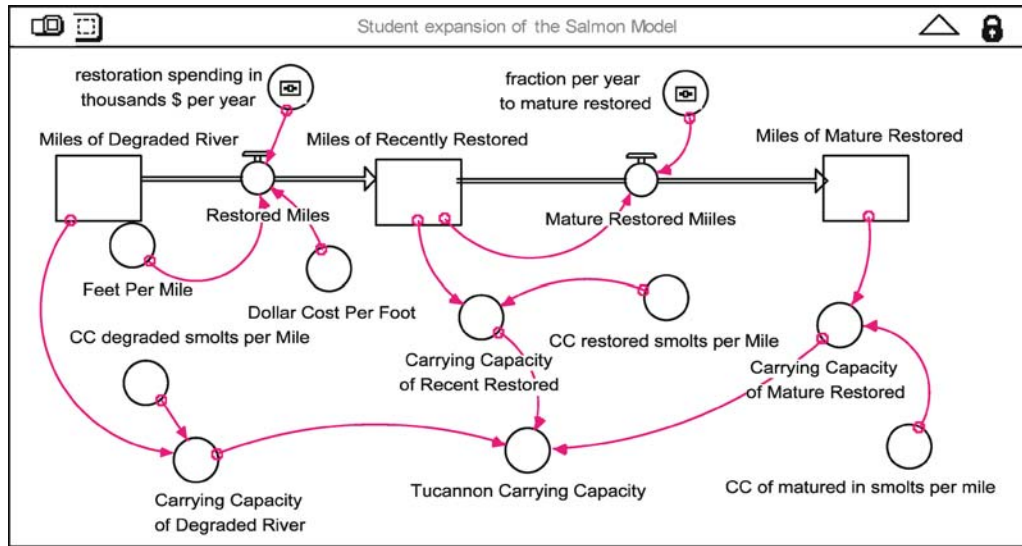
Figure 8 shows a version of the model to encourage student experimentation with harvesting policies. The information fields instruct the students to work in groups of three with one student playing the role of “the harvest manager”. The harvest manager’s goal is to achieve a large, sustainable harvest through control of the harvest fraction. The other students are given control of the parameters that describe conditions on the Snake and Columbia and in the Tucannon watershed. These students are encouraged to make major and unpredictable changes to test the instincts of the harvest manager.

Models designed for highly interactive simulations of this kind are sometimes called “management flight simulators” because they serve the same function as actual flight simulators. With a pilot simulator, the trainee takes the controls of an electro-mechanical model and tests his instincts for managing the simulated airplane under difficult conditions. The Tucannon harvesting model provides a similar opportunity for environmental students. They can learn the challenge of managing open access fisheries that are vulnerable to over harvesting and the tragedy

of the commons [12]. In this particular exercise, students learn that they can achieve a sustainable harvest under a wide variety of difficult and unpredictable conditions. The key to sustainability is harvest manager’s freedom to change the harvest fraction in response to recent trends in number of returning adults. This is an important finding for fishery management because it reveals that the population dynamics are not the main obstacle to sustainability. Rather, unsustainable harvesting is more likely to occur when the managers find it difficult to change the harvest fraction in response to recent trends. This is the fundamental challenge of an open-access fishery.

The salmon model is a system dynamics version of the type of modeling commonly performed by population biologists. System dynamics adds clarity and ease of experimentation compared to these models. It also provides a launching point for model expansions that can go beyond population biology. Figure 9 shows an example. This is a student expansion to change the carrying capacity from a user input to a variable that responds to the user’s river restoration strategy. The student was trained in geomorphology and was an expert on restoring degraded rivers in the west. The Tucannon began the simulation with 25 miles of river in degraded condition and the remaining 25 miles in a mature, fully restored river with a much higher carrying capacity. The new model per-

shocks are external to the boundary of this model, so one is reminded of Coyle’s definition of system dynamics. That is, the model helps us understand how the salmon population could withstand the shocks which fall upon it from the out-side world.



System Dynamics Models of Environment, Energy and Climate Change, Figure 9  
Student addition to simulate river restoration

mits one to experiment with the timing of river restoration spending and to learn the impact on the management of the salmon fishery.

The student's model provides another example of interdisciplinary modeling that aids our understanding of environmental systems. In this particular case, the modeling of river restoration is normally the domain of the geomorphologist. The model of the salmon population is the domain of the population biologist. Their work is often conducted separately, and their models are seldom connected. This is unfortunate as the experts working in their separate domains miss out on the insights that arise when two perspectives are combined within a single model. In the student's case, surprising insights emerged when the combined model was used to study the economic value of the harvesting that could be sustained in the decades following the restoration of the river. To the student's surprise, the new harvesting could "pay back" the entire cost of the river restoration in less than a decade.

### Models of Climate Change

Scientists use a variety of models to keep track of the greenhouse gasses and their impact on the climate. Some of the models combine simulations of the atmosphere, soils, biomass and ocean response to anthropogenic emissions. The more developed models include CO<sub>2</sub>, methane, nitrous oxides and other greenhouse gas (GHG) emissions and their changing concentrations in the atmosphere. Claussen [2] classifies climate models as simple, interme-

diate and comprehensive. The simple models are sometimes called "box models" since they represent the storage in the system by highly aggregated stocks. The parameters are usually selected to match the results from more complicated models. The simple models can be simulated faster on the computer, and the results are easier to interpret. This makes them valuable for sensitivity studies and in scenario analysis [13].

The comprehensive models are maintained by large research centers, such as the Hadley Center in the UK. The term "comprehensive" refers to the goal of capturing all the important processes and simulating them in a highly detailed manner. The models are sometimes called GCMs (general circulation models). They can be used to describe circulation in the atmosphere or the ocean. Some simulate both the ocean and atmospheric circulation in a simultaneous, interacting fashion. They are said to be coupled general circulation models (CGCMs) and are considered to be the "most comprehensive" of the models available [2]. They are particularly useful when a high spatial resolution is required. However, a disadvantage of the CGCMs is that only a limited number of multi-decadal experiments can be performed even when using the most powerful computers.

Intermediate models help scientists bridge the gap between the simple and the comprehensive models. Claussen [2] describes eleven models of intermediate complexity. These models aim to "preserve the geographic integrity of the Earth system" while still providing the opportunity for multiple simulations to "explore the param-

eter space with some completeness. Thus, they are more suitable for assessing uncertainty”. Figure 10 characterizes the different categories of models based on their relative emphasis on:

- number of processes (right axis)
- detailed treatment of the each process (left axis), and the
- extent of integration among the different processes (top axis).

Regardless of the methodology, climate modeling teams must make some judgments on where to concentrate their attention. No model can achieve maximum performance along all three dimensions. (Figure 10 uses the dashed lines to draw our attention to the impossible task of doing every thing within a single model.)

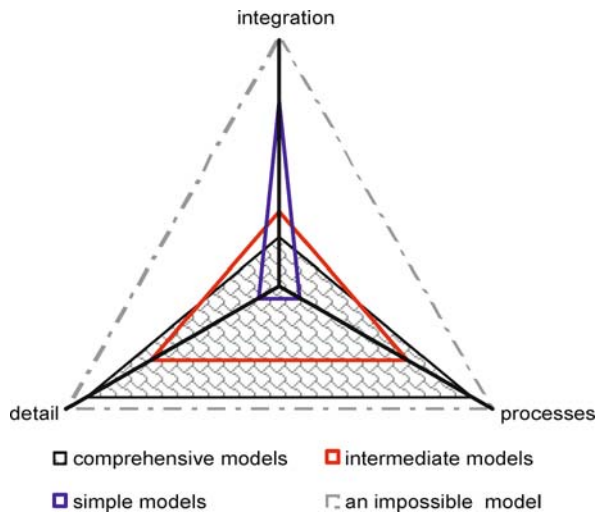
The comprehensive models strive to simulate as many processes as possible with a high degree of detail. This approach provides greater realism, but the models often fail to simulate the key feedback loops the link that atmospheric system with the terrestrial and oceanic systems. (An example is the feedback between CO<sub>2</sub> emissions, temperatures and the decomposition of soil carbon. If higher temperatures lead to accelerated decomposition, the soils could change from a net sink to a net source of carbon [15].) The simple models sacrifice detail and the number of processes in order to focus on the feedback effects between the processes. Using Claussen’s terminology, one would say that such models aim for a high degree of “integration”. However, the increased integration is achieved by limiting the number of processes and the degree of detail in representing each of the processes.

System dynamics has been used in a few applications to climate change. These applications fit in the category of simple models whose goal is to provide a highly integrated representation of the system. Two examples are described here; both deal with the complexities of the global carbon cycle.

### System Dynamics Models of the Carbon Cycle

Figures 11 and 12 depict the global carbon cycle. Figure 11 shows the carbon flows in a visual manner. Figure 12 uses the Vensim stock and flow icons to summarize carbon storage and flux in the current system. The storage is measured in GT, gigatons of carbon, (where carbon is the C in CO<sub>2</sub>). The flows are in GT/year of carbon with values rounded off for clarity.

The left side of Fig. 12 shows the flows to the terrestrial system. The primary production removes 121 GT/yr from the atmosphere. This outflow exceeds the return flows by



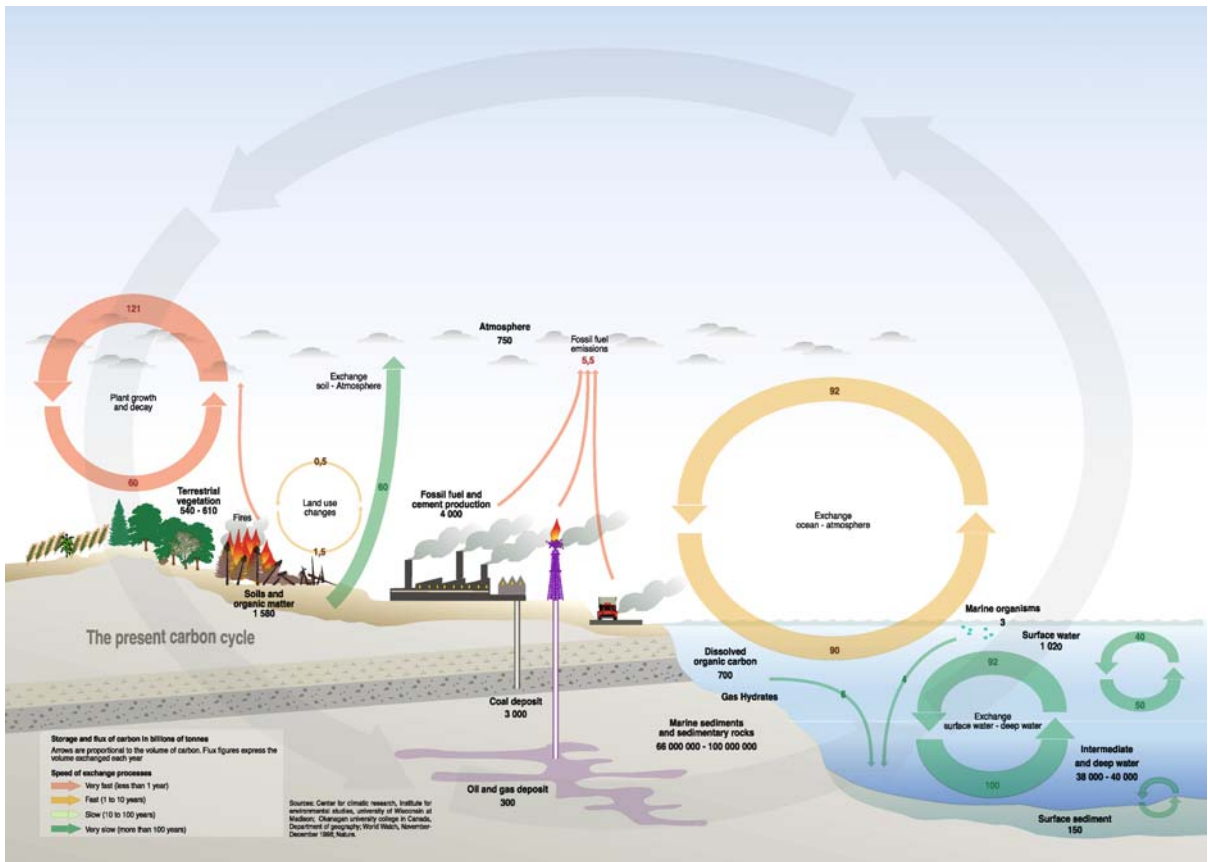
System Dynamics Models of Environment, Energy and Climate Change, Figure 10

Classification of climate models

1 GT/year. This imbalance suggests that around 1 GT of carbon is added to the stocks of biomass and soil each year. So the carbon stored in the terrestrial system would grow over time (perhaps due to extensive reforestation of previously cleared land.) The right side of Fig. 12 shows the flows from the atmosphere to the ocean. The CO<sub>2</sub> dissolved in the ocean each year exceeds the annual release back to the atmosphere by 2 GT. The total, net-flow out of the atmosphere is 3 GT/year which means that natural processes are acting to negate approximately half of the current anthropogenic load.

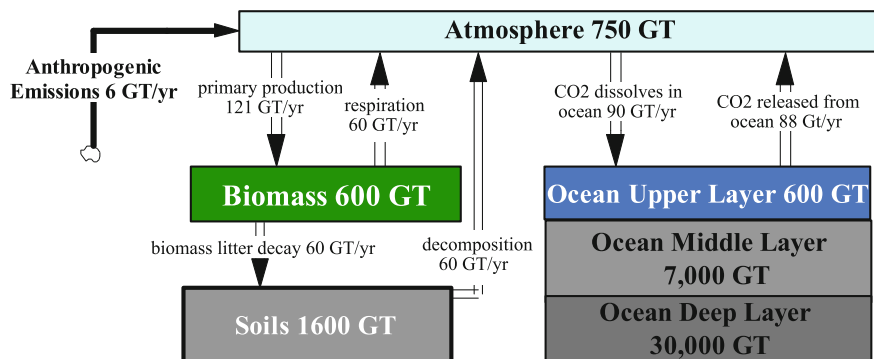
As the use of fossil fuels grows over time, the anthropogenic load will increase. But scientists do not think that natural processes can continue to negate 50% of an ever increasing anthropogenic load. On the terrestrial side of the system, there are limits on the net flow associated with reforestation of previously cleared land. And there are limits to the carbon sequestration in plants and soils due to nitrogen constraints. On the ocean side of the system, the current absorption of 2 GT/year is already sufficiently high to disrupt the chemistry of the ocean’s upper layer. Higher CO<sub>2</sub> can reduce the concentration of carbonate, the ocean’s main buffering agent, thus affecting the ocean’s ability to absorb CO<sub>2</sub> over long time periods.

Almost of the intermediate and comprehensive climate models may be used to estimate CO<sub>2</sub> accumulation in the atmosphere in the future. For this article, it is useful to draw on the mean estimate published in *Climatic Change* by Webster [23]. He used the climate model developed at the Massachusetts Institute of Technology, one of the eleven models of “intermediate complexity” in the



System Dynamics Models of Environment, Energy and Climate Change, Figure 11

The global carbon cycle. (Source: United Nations Environmental Program (UNEP) <http://www.unep.org/>)



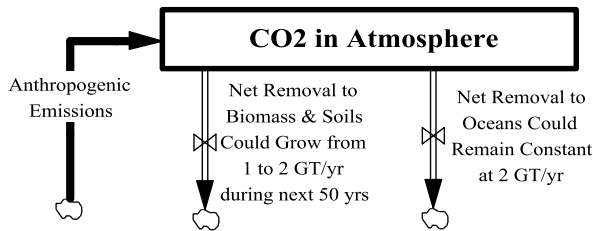
System Dynamics Models of Environment, Energy and Climate Change, Figure 12

Diagram of the stocks and flows in the carbon cycle

review by Claussen [2]. The model began the simulation in the year 2000 with an atmospheric CO<sub>2</sub> concentration of 350 parts per million (ppm). (This concentration corresponds to around 750 GT of carbon in the atmosphere.) The mean projection assumed that anthropogenic emissions would grow to around 19 GT/year by 2100. The

mean projection of atmospheric CO<sub>2</sub> was around 700 ppm by 2100. The amount of CO<sub>2</sub> in the atmosphere would be twice as high at the end of the century.

Figure 13 shows the simplest possible model to explain the doubling of atmospheric CO<sub>2</sub>. The stock accumulates the effect of three flows, each of which is specified by the



**System Dynamics Models of Environment, Energy and Climate Change, Figure 13**

**Simple model to understand accumulation of CO<sub>2</sub> in the atmosphere**

user. Anthropogenic emissions are set to match Webster's assumption. They grow to 19 GT/year by the end of the century. Net removal to oceans is assumed to remain constant at 2 GT/year for the reasons given previously. Net removal to biomass and soils is then subject to experimentation to allow this simple model to match Webster's results. A close match is provided if the net removal increases from 1 to 2 GT/year during the first half of the century and then remains at 2 GT/year for the next fifty years. With these assumptions, the CO<sub>2</sub> in the atmosphere would double from 750 to 1500 GT during the century. This means that the atmospheric concentration would double from 350 to 700 ppm, the same result published by Webster [23].

The model in Fig. 13 is no more than an accumulator. This is the simplest of possible models to add insight on the dynamics of CO<sub>2</sub> accumulation in the atmosphere. It includes a single stock and only three flows, with all of the flows specified by the user. There are no feedback relationships which are normally at the core of system dynamics models. This extreme simplification is intended to make the point that simple models may provide perspective on the dynamics of a system. In this case, a simple accumulator can teach one about the sluggish response of atmospheric CO<sub>2</sub> in the wake of reductions in the anthropogenic emissions. As an example, suppose carbon policies were to succeed in cutting global emissions dramatically in the year 2050. By this year, emissions would have reached 10 GT/yr, so the supposed policy would reduce emissions to 5 GT/yr. What might then happen to CO<sub>2</sub> concentrations in the atmosphere for the remainder of the century? Experiments with highly educated adults [21] suggest that some subjects would answer this question with "pattern matching" reasoning. For example, if emissions are cut in half, it might make sense that CO<sub>2</sub> concentrations would be cut in half as well. But pattern matching leads one astray since the accumulation of CO<sub>2</sub> in the atmosphere responds to the total effect of the flows in Fig. 13. Were anthropogenic emissions to be reduced to 5 GT/year

and net removals were to remain at 4 GT/year, the CO<sub>2</sub> concentration would continue to grow, and atmospheric CO<sub>2</sub> would reach 470 ppm by the end of the century.

The model in Fig. 13 is an extreme example to make a point about the usefulness of simple models. The next example is by Fiddaman [6]. It was selected as illustrative of the type of model that would emerge after a system dynamics study. Figure 14 shows the view of the carbon cycle, one of 30 views in the model. The model simulates the climate system within a larger system that includes growth in human population, growth in the economy, and changes in the production of energy. The model was organized conceptually as nine interacting sectors with a high degree of coupling between the energy, economic and the climate sectors.

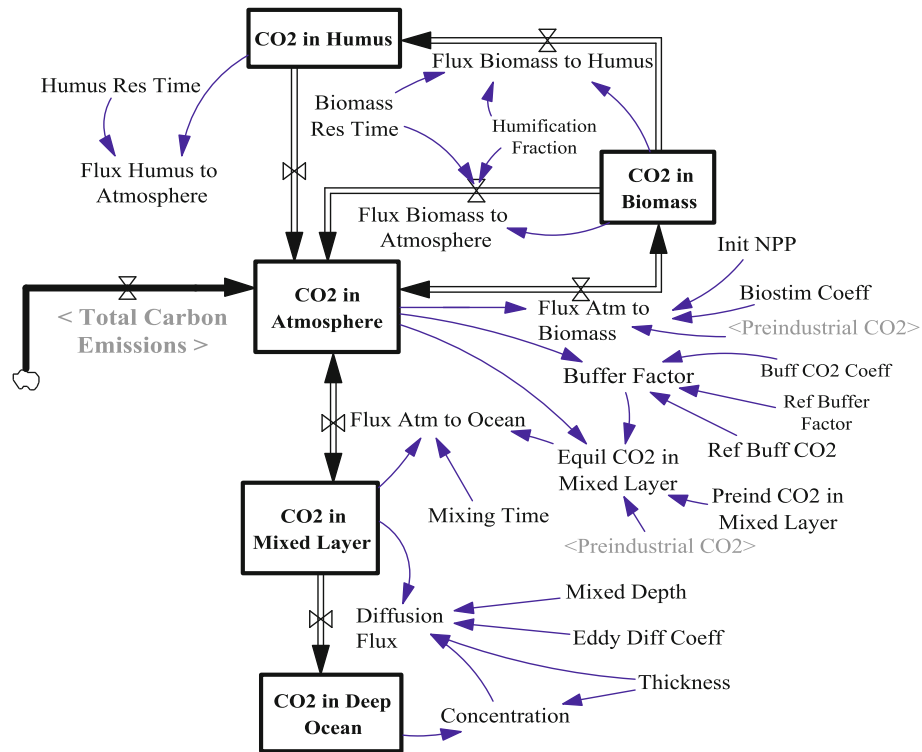
Fiddaman focused on policy making, particularly the best way to put a price on carbon. In the current debate, this question comes down to a choice between a carbon tax and a carbon market. His simulations add support to those who argue that the carbon tax is the preferred method of putting a price on carbon. The simulations also provide another example of the usefulness of system dynamics models that cross disciplinary boundaries. By representing the economy, the energy system and the climate system within a single, tightly coupled model, he provides another example of the power of system dynamics to promote interdisciplinary exploration of complex problems.

System dynamics has also been applied to a wide variety of energy problems [1,7]. Indeed, a key word frequency count in 2004 revealed nearly 400 energy entries in the System dynamics bibliography [11]. Many of these applications deal with the electric power industry, and I have selected two electric studies to illustrate the usefulness of the approach. The first involves the regulatory and financial challenges of the investor owned electric utilities in the United States.

### Lessons from the Regulated Power Industry in the 1970s

The 1970s was a difficult decade for the regulated power companies in the United States. The price of oil and gas was increasing rapidly, and the power companies were frequently calling on their regulators to increase retail rates to cover the growing cost of fuel. The demand for electricity had been growing rapidly during previous decades, often at 7%/year. At this rate, the demand doubled every decade, and the power companies faced the challenge of doubling the amount of generating capacity to ensure that demand would be satisfied. The power companies dealt with this challenge in previous decades by building ever





System Dynamics Models of Environment, Energy and Climate Change, Figure 14  
Representation of the carbon cycle in the model by Fiddaman [6]

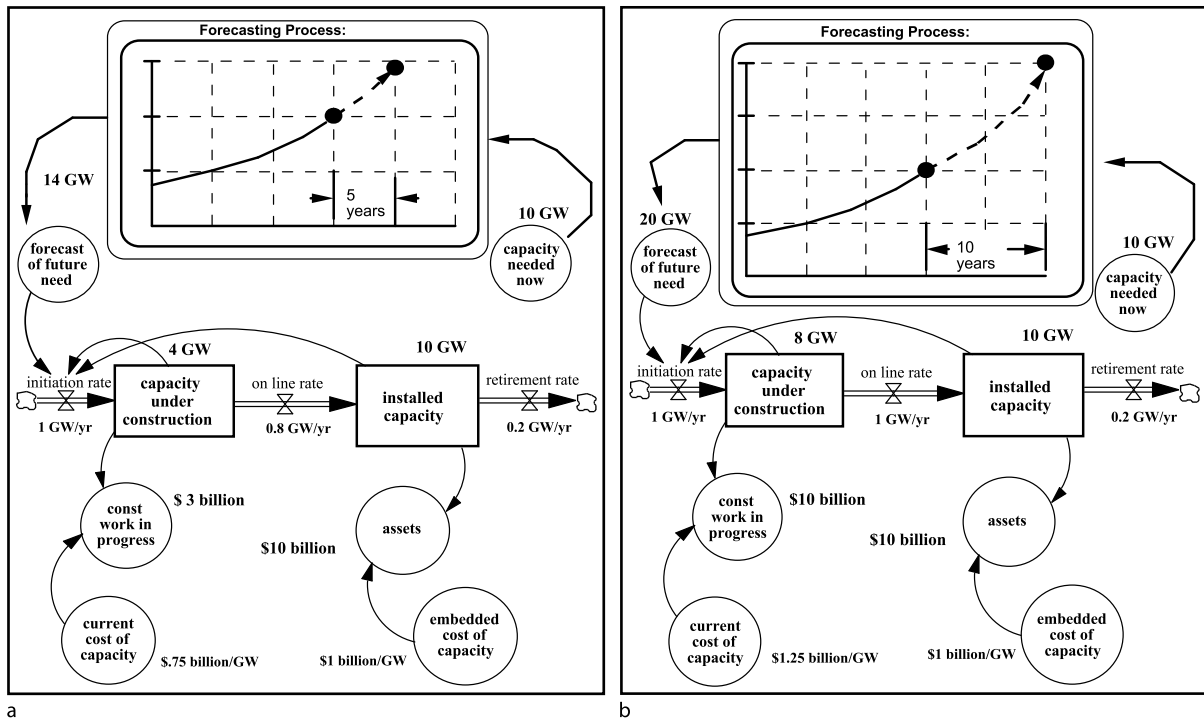
larger power plants (whose unit construction costs declined due to economies of scale). But the economies of scale were exhausted by the 1970s, and the power companies found themselves with less internal funds and poor financial indicators. Utilities worried that the construction of new power plants would not keep pace with demand, and the newspapers warned of curtailments and blackouts.

Figure 15 puts the financial problems in perspective by showing the forecasting, planning and construction processes. The side by side charts allows one to compare the difficult conditions of the 1970s with conditions in previous decades. Figure 15a shows the situation in the 1950s and 1960s. Construction lead times were around 5 years, so forecasts would extend 5 years into the future. Given the costs at the time, the power company would need to finance \$3 billion in construction. This was a substantial, but manageable task for a company with \$10 billion in assets.

Figure 15b shows the dramatic change in the 1970s. Construction lead times had grown to around 10 years, and construction costs had increased as well. The power company faced the challenge of financing \$10 billion in construction with an asset base of \$10 billion. The utility executives turned to the regulators for help. They asked

for higher electricity rates in order to increase annual revenues and improve their ability to attract external financing. The regulators responded with substantial rate increases, but they began to wonder whether further rate increases would pose a problem with consumer demand. If consumers were to lower electricity consumption, the utility would have less sales and less revenues. The executives might then be forced to request another round of rate increases. Regulators wondered if they were setting loose a “death spiral” of ever increasing rates, declining sales and inadequate financing.

Figure 16 puts the problem in perspective by showing the consumer response to higher electricity rates along side of the other key feedback loops in the system. Higher electricity rates do pose the problem which came to be called “the death spiral”. But the death spiral does not act in isolation. Figure 16 reminds us that higher rates lead to lower consumption and to a subsequent reduction in the demand forecast and in construction. After delays for the new power plants to come on line, the power companies experiences a reduction in its “rate base” and the “allowed revenues”. When the causal relationships are traced around the outer loop, one sees a negative feedback loop that could act to stabilize the situation. The prob-



System Dynamics Models of Environment, Energy and Climate Change, Figure 15

a The electric utility's financial challenge during the 1950s and 1960s. b The electric utility's financial challenge during the 1970s

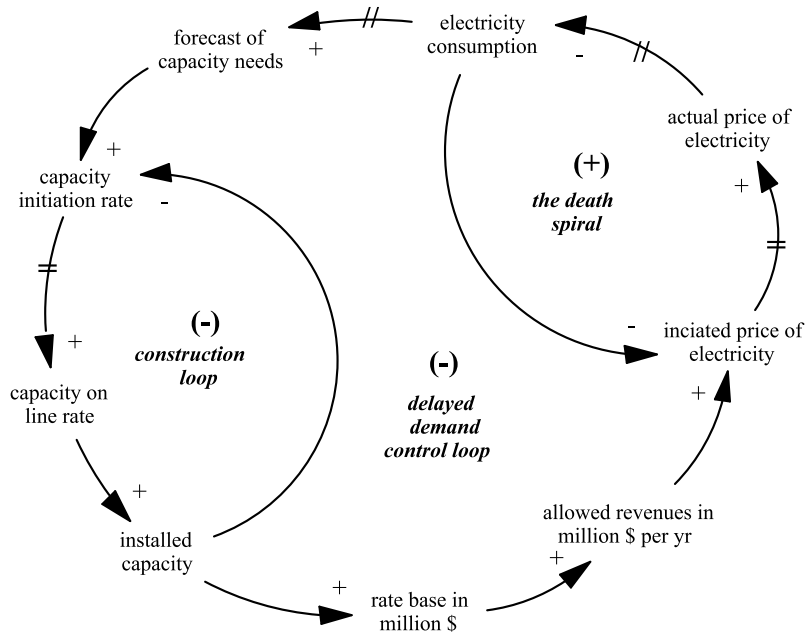
lem, however, is that the delays around the outer loop are substantially longer than the delay for the death spiral.

The utility companies financial challenge was the subject of several system dynamics studies in the 1970s and 1980s [7]. The studies revealed that the downward spiral could pose difficult problems, especially if consumers reacted quickly while utilities were stuck with long-lead time, capital intensive power plants under construction. The studies showed that utility executives needed to do more than rely on regulators to grant rate increases; they needed to take steps on their own to soften the impact of the death spiral. The best strategy was to shift the investments to technologies with shorter lead times. (As an example, a power company in coal region would do better to switch from large to smaller coal plants because of the small plants' shorter lead time.) The studies also revealed that the company's financial situation would improve markedly with slower growth in demand. By the late 1970s and early 1980s, many power companies began to provide direct financial incentives to their customers to slow the growth in demand. System dynamics studies showed that the company-sponsored efficiency programs would be beneficial to the both the customers (lower elec-

tric bills) and to the power companies (improved financial performance).

An essential feature of the utility modeling was the inclusion of power operations along side of consumer behavior, company forecasting, power plant construction, regulatory decision making and company financing. This interdisciplinary approach is common within the system dynamics community because practitioners believe that insights will emerge from simulating the key feedback loops. (This belief leads one to follow the cause and effect connections around the key loops regardless of the disciplinary boundaries that are crossed along the way.) This approach contrasts strongly with the customary modeling framework of large power companies who were not familiar with system dynamics. Their approach was to assign models to different departments (i. e., operations, accounting and forecasting) and string the models together to provide a view of the entire corporation over the long-term planning interval.

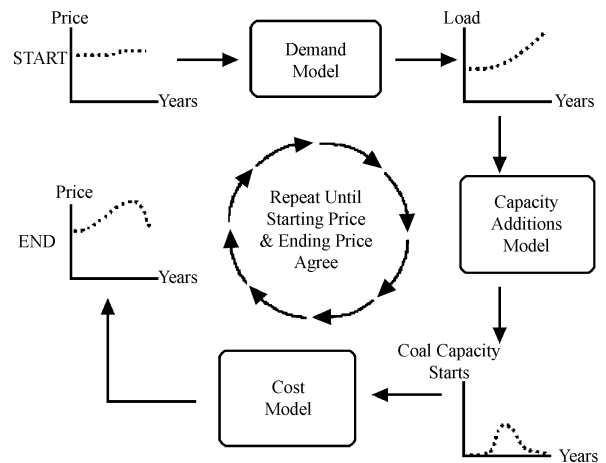
Figure 17 shows what can happen when models within separate departments are strung together. A large corporation might use 30 models, but this diagram makes the point by describing three models. The analysis would begin with an assumption on future electricity prices over the



**System Dynamics Models of Environment, Energy and Climate Change, Figure 16**  
Key feedbacks and delays faced by power companies in the 1970s

20-year interval. These are needed to prepare a forecast of the growth in electricity load. The forecast is then given to the planning department which may run a variety of models to select the number power plants to construct in the future. The construction results are then handed to the accounting and rate making departments to prepare a forecast of electricity prices. When the company finally completes the many calculations, the prices that emerge may not agree with the prices that were assumed at the start. The company must then choose whether to ignore the contradiction or to repeat the entire process with a new estimate of the prices at the top of the diagram. This was not an easy choice. Ignoring the price discrepancy was problematic because it was equivalent to ignoring the “death spiral,” one of the foremost problems of the 1970s. Repeating the analysis was also problematic. The new round of calculations would be time consuming, and there was no guarantee that consistent results would be obtained at the end of the next iteration.

The power companies’ dilemma from the 1970s is described here to make an important point about the usefulness of system dynamics. System dynamics modeling is ideally suited for the analysis of dynamic problems that require a feedback perspective. The method allows one to “close the loop”, as long as one is willing to cross the necessarily disciplinary boundaries. In contrast, other modeling methods are likely to be extremely time consuming or fall



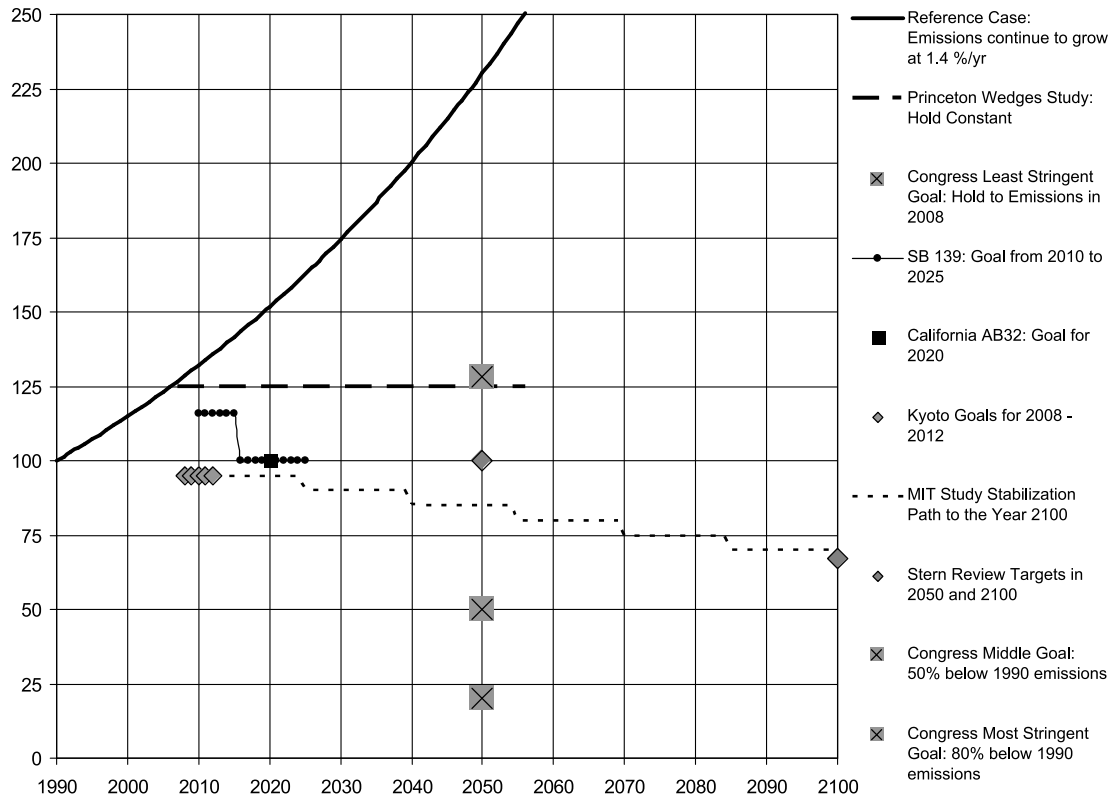
**System Dynamics Models of Environment, Energy and Climate Change, Figure 17**

The iterative approach often used by large power companies in the 1970s

short in simulating the key feedbacks that tie the system together.

### Simulating the Power Industry Response to a Carbon Market

The world is getting warmer, both in the atmosphere and in the oceans. The clearest and most emphatic description



System Dynamics Models of Environment, Energy and Climate Change, Figure 18

Comparison of goals for emissions (100 on the vertical axis represent emissions in the year 1990)

of global warming was issued by the intergovernmental panel on climate change (IPCC) in February of 2007. Their summary for policymakers (p. 4 in [14]) reported that the “Warming of the climate system is unequivocal, as is now evident from observations of increases in global average air and ocean temperatures, widespread melting of snow and ice and rising global mean sea level”. The IPCC concluded that “most of the observed increase is very likely due to the observed increase in anthropogenic greenhouse gas concentrations”. As a consequence of the IPCC and other warnings, policymakers around the world are calling for massive reductions in CO<sub>2</sub> and other greenhouse gas (GHG) emissions to reduce the risks of global warming.

Figure 18 summarizes some of the targets for emission reductions that have been adopted or proposed around the world. In many cases, the targets are specified relative to a country’s emissions in the year 1990. So, for ease of comparison, the chart uses 100 to denote emissions in the year 1990. Emissions have been growing at around 1.4%/year. The upward curve shows the future emissions if this trend continues: emissions would reach 200 by 2040 and 400 by 2090. The chart shows the great differences in the strin-

gency of the targets. Some call for holding emissions constant; others call for dramatic reductions over time. Some targets apply to the next two decades; many extend to the year 2050; and some extend to the year 2100. However, when compared to the upward trend, all targets require major reductions relative to business as usual.

The targets from the Kyoto treaty are probably the best known of the goals in Fig. 18. The treaty became effective in February of 2005 and called for the Annex I countries to reduce emissions, on average, by 5% below 1990 emissions by the year 2008 and to maintain this limit through 2012. The extension of the Kyoto protocol beyond 2012 is the subject of ongoing discussions. The solid line from 2010 to 2050 represents the “stabilization path” used in the climate modeling by Webster [23]. The limit on emissions was imposed in modeling calculations designed to stabilize atmospheric CO<sub>2</sub> at 550 ppmv or lower. The scenario assumed that the Kyoto emissions caps are adopted by all countries by 2010. The policy assumed that the caps would be extended and then further lowered by 5% every 15 years. By the end of the century, the emissions would be 35% below the value in 1990.



System Dynamics Models of Environment, Energy and Climate Change, Figure 19

Map of the western electricity system

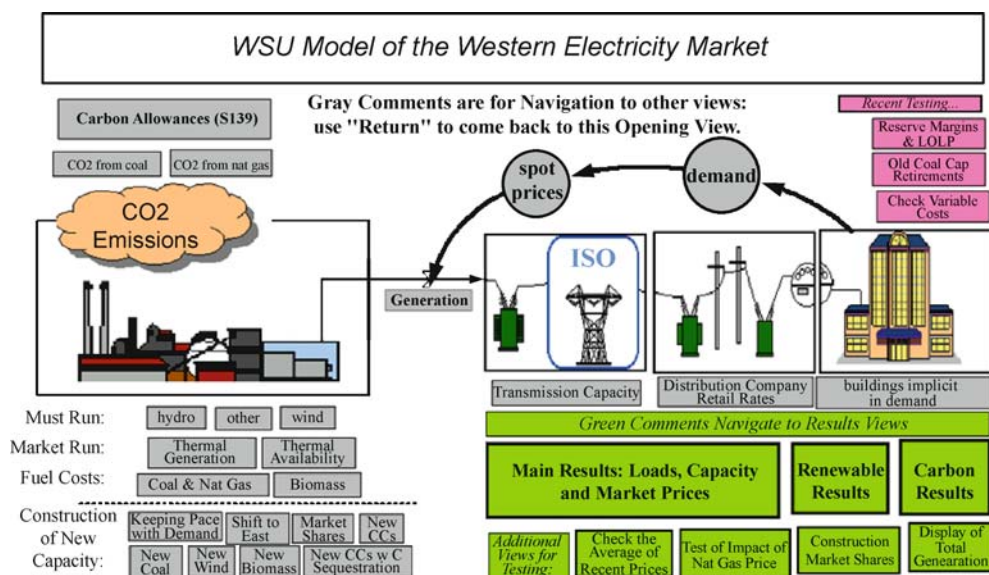
This article concentrates on Senate Bill 139, The Climate Stewardship Act of 2003. Figure 19 shows the S139 targets over the interval from 2010 to 2025. The bill called for an initial cap on emissions from 2010 to 2016. The cap would be reduced to a more challenging level in 2016, when the goal was to limit emissions to no more than the emissions from 1990. S139 was introduced by Senators McCain and Lieberman in January of 2003. It did not pass, but it was the subject of several studies including a highly detailed study by the Energy Information Administration [5]. The EIA used a wide variety of models to

search for the carbon market prices that would induce industries to lower emissions to come into compliance with the cap. The carbon prices were estimated at \$22 per metric ton of CO<sub>2</sub> when the market was to open in 2010. They were projected to grow to \$60 by the year 2025.

The EIA study showed that the electric power sector would lead the way in reducing emissions. By the year 2025, power sector emissions would be reduced 75% below the reference case. This reduction was far beyond the reductions to be achieved by other sectors of the economy. This dramatic response was possible given the large use of coal in power generation and the power industry's wide range of choices for cleaner generation.

A system dynamics study of S139 was conducted at Washington State University (WSU) to learn if S139 could lead to similar reductions in the west. Electricity generation in the western system is provided in a large, interconnected power system shown in Fig. 19. This region has considerably more hydro resources, and it makes less use of coal-fired generation than the nation as a whole. The goal was to learn if dramatic reductions in CO<sub>2</sub> emissions could be possible in the west and to learn if they could be achieved with generating technologies that are commercially available today.

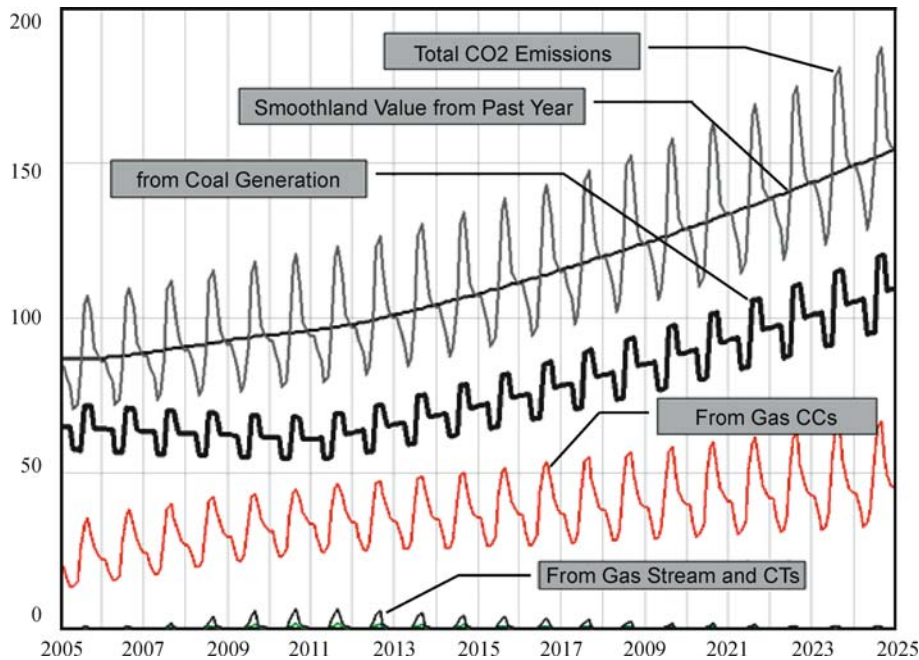
The opening view of the WSU model is shown in Fig. 20. The model deals with generation, transmission and distribution to end use customers, with price feedback on the demand for electricity. The model is much larger than the textbook models described earlier in this article. Fifty



System Dynamics Models of Environment, Energy and Climate Change, Figure 20

Opening view of the model of the western electricity system





System Dynamics Models of Environment, Energy and Climate Change, Figure 21

Annual emissions in a base case simulation (annual emissions are in million metric tons of carbon)

views are required to show the all the diagrams and the simulation results. The opening view serves as a central hub to connect with all the other views.

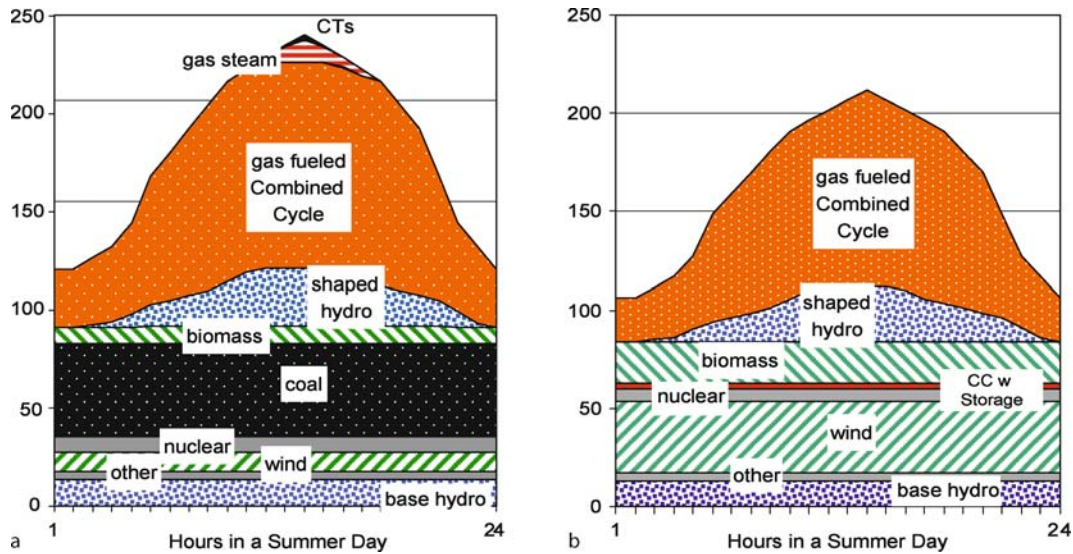
The opening view uses Vensim's comment icons to draw attention to the CO<sub>2</sub> emissions in the model. The emissions arise mainly from coal-fired power plants, as shown in Fig. 21. A smaller, but still significant fraction of the emissions is caused by burning natural gas in combined cycle power plants. Total emissions vary with the seasons of the year, with the peak normally appearing in the summer when almost all of the fossil-fueled plants are needed to satisfy peak demand. The base case shows annual emissions growing by over 75% by the year 2025.

A major challenge for the system dynamics model is representing power flows across a transmission grid. Finding the flows on each transmission line and the prices in each area is difficult with the standard tools of system dynamics. It simply doesn't make sense to represent the power flows with a combination of stocks, flows and feedback processes to explain the flows. It makes more sense to calculate the flows and prices using traditional power systems methods, as explained by Dimitrovski [4]. The power flows were estimated using an algebraic approach which power engineers label as a reduced version of a direct-current optimal power flow calculation. The solution to the algebraic constraints were developed with the Matlab software and then transferred to user-defined functions to op-

erate within the Vensim software. The Vensim simulations were set to run over twenty years with time in months. (A typical simulation required 240 months with changes during a typical day handled by carrying along separate calculations for each of 24 h in a typical day.) These are extensive calculations compared to many system dynamics models, so there was concern that we would lose the rapid simulation speed that helps to promote interactive exploration and model testing. The important methodological accomplishment of this project was the inclusion of network and hourly results within a long-term model without losing the rapid simulation response that encourages users to experiment with the model.

One of the model experiments called for a new simulation with carbon prices set to follow the \$20 to \$60 trajectory projected by the EIA for \$139. These prices were specified as a user input, and the model responded with a change in both short-term operations and long-term investments. The important result was a 75% reduction in CO<sub>2</sub> emissions by the end of the simulation. This dramatic reduction corresponds almost exactly to the EIA estimate of CO<sub>2</sub> reduction for the power industry in the entire US.

Figure 22 helps one understand how CO<sub>2</sub> emissions could be reduced by such a large amount. These diagrams show the operation of generating units across the Western US and Canada for a typical day in the summer of the final year of the simulation. Figure 22a shows the reference



System Dynamics Models of Environment, Energy and Climate Change, Figure 22

**a** Projected generation for a peak summer day in 2024 in the reference case. **b** Projected generation for a peak summer day in 2024 in the S139 case

case; Figure 22b shows the case with S139. The side by side comparison helps one visualize the change in system operation. A comparison of the peak loads shows that the demand for electricity would be reduced. The reduction is 9%, which is due entirely to the consumers' reaction to higher retail electric prices over time.

Figure 22b shows large contributions from wind and biomass generation. Wind generation is carbon free, and biomass generation is judged to be carbon neutral, so these generating units make an important contribution by the end of the simulation. Both of these generating technologies are competitive with today's fuel prices and tax credits. The model includes combined cycle gas generation equipped with carbon capture and storage, a technology that is not commercially available today. The model assumes that advances in carbon sequestration over the next two decades would allow this technology to capture a small share of investment near the end of the simulation. By the year 2025, the combined cycle plants with sequestration equipment would provide 2% of the generation.

The most important observation from Fig. 22 is the complete elimination of coal-fired generation in the S139 case. Coal-fired units are shown to operate in a base load mode in the reference simulation. They provide around 28% of the annual generation, but they account for around two-thirds of the CO<sub>2</sub> emissions in the western system. The carbon prices from S139 make investment in new coal-fired capacity unprofitable at the very start of the simulated market in 2010. As the carbon prices increase, util-

ities to cut back on coal-fired generation and compensate with increased generation from gas-fired CC capacity. In the simulations reported here, this fuel switching would push the coal units into the difficult position of operating fewer and fewer hours in a day. Eventually this short duration operation is no longer feasible, and coal generation is eliminated completely by the end of the simulation.

The WSU study of the western electric system was selected as the concluding example because of its novel treatment of network flows inside a system dynamics model [4]. The model is also interesting for its treatment of daily price changes within a long-term model. (Such changes are important to the simulation of revenues in the wholesale market.) From a policy perspective, the study confirms previous modeling of the pivotal role of the electric power industry in responding to carbon markets. The study indicated that the western electricity system could achieve dramatic reductions in CO<sub>2</sub> emissions within 15 years after the opening of a carbon market, and it could do so with technologies that are commercially available today [8].

### Conditions for Effective Interdisciplinary Modeling

All of the applications demonstrate the usefulness of system dynamics in promoting interdisciplinary modeling. The article concludes with comments on the level of effort and the conditions needed for effective, interdisciplinary modeling.

The examples in this article differ substantially in the level of effort required, from several weeks for the classroom examples to several years for the energy studies. The textbook examples involved student expansions of models of Mono Lake and the Tucannon salmon. The expansions were completed by undergraduate students in projects lasting two or three weeks. The key was the students' previous education (classes from many different departments) and their receptiveness to an interdisciplinary approach.

Fiddaman's model of the climate and energy system [6] was a more ambitious exercise, requiring several years of effort as part of his doctoral research. Bringing multi-year interdisciplinary modeling projects to a successful conclusion requires one to invest the time to master several disciplines and to maintain a belief that there are potential insights at the end of the effort.

The electric power industry examples were also ambitious projects that required several years of effort. The modeling of the western electricity system was a four-year project with support from the National Science Foundation. The long research period was crucial for it allowed the researchers from power systems engineering, system dynamics and environmental science to take the time to learn from one another. The modeling of the electric company problems in the 1970s was also spread over several years of effort. The success of this modeling was aided by utility planners, managers and modelers who were looking for a systems view of their agency and its problems. They saw system dynamics as a way to tie existing ideas together within an integrated portrayal of their system. Their existing ideas were implemented in models maintained by separate functional areas (i. e., forecasting, accounting, operations). The existing models often provided a foundation for the system dynamics models (i. e., in the same way that the comprehensive climate models in Fig. 11 provide support for the development of the more integrated models). The key to effective, interdisciplinary modeling within such large organizations is support from a client with a strong interest in learning and with managerial responsibility for the larger system.

## Future Directions

This article concludes with future directions for system dynamics applications to climate change. People often talk of mitigation and adaptation. Mitigation refers to the challenge of lowering greenhouse gas emissions to avoid dangerous anthropogenic interference with the climate system. Adaptation refers to the challenge of living in a changing world.

**Mitigation:** The challenge of lowering CO<sub>2</sub> and other GHG emissions is the fundamental challenge of the coming century. The next two decades will probably see various forms of carbon markets, and system dynamics can aid in learning about market design. It is important that we learn how to make these markets work well. And if they don't work well, it's important to speed the transition to a carbon tax policy with better prospects for success. System dynamics can aid in learning about markets, especially if it is coupled with simulating gaming to allow market participants and regulators to "experience" and better understand market dynamics.

**Adaptation:** The world will continue to warm, and sea levels will continue to rise. These trends will dominate the first half of this century even with major reductions in CO<sub>2</sub> emissions. These and other climate changes will bring a wide variety of problems for management of water resource, public health planning, control of invasive species, preservation of endangered species, control of wildfire, and coastal zone management, just to name a few. Our understanding of the adaptation challenges can be improved through system dynamics modeling. The prospects for insight are best if the models provide an interdisciplinary perspective on adapting to a changing world.

## Bibliography

### Primary Literature

1. Bunn D, Larsen E (1997) Systems modelling for energy policy. Wiley, Chichester
2. Claussen M et al (2002) Earth system models of intermediate complexity: closing the gap in the spectrum of climate system models. *Climate Dyn* 18:579–586
3. Coyle G (1977) Management system dynamics. Wiley, Chichester
4. Dimitrovski A, Ford A, Tomsovic K (2007) An interdisciplinary approach to long term modeling for power system expansion. *Int J Crit Infrastruct* 3(1–2):235–264
5. EIA (2003) United States Department of Energy, Energy Information Administration, Analysis of S139, the Climate Stewardship Act of 2003
6. Fiddaman T (2002) Exploring policy options with a behavioral climate-economy model. *Syst Dyn Rev* 18(2):243–264
7. Ford A (1999) Modeling the environment. Island Press, Washington
8. Ford A (2008) Simulation scenarios for rapid reduction in carbon dioxide emissions in the western electricity system. *Energy Policy* 36:443–455
9. Forrester J (1961) Industrial dynamics. Pegasus Communications
10. Forrester J (2000) From the ranch to system dynamics: An autobiography, in management laureates. JAI Press
11. Ford A, Cavana R (eds) (2004) Special Issue of the *Syst Dyn Rev*
12. Hardin G (1968) The tragedy of the commons. *Science* 162:1243–1248

13. IPCC (1997) An introduction to simple climate models used in the IPCC second assessment report. ISBN 92-9169-101-1
14. IPCC (2007) Climate change 2007: The physical science basis, summary for policymakers. [www.ipcc.ch/](http://www.ipcc.ch/)
15. Kump L (2002) Reducing uncertainty about carbon dioxide as a climate driver. *Nature* 419:188–190
16. Meadows DH, Meadows DL, Randers J, Behrens W (1972) The limits to growth. Universe Books
17. Morecroft J (2007) Strategic modelling and business dynamics. Wiley, Chichester
18. Richardson J, Pugh A (1981) Introduction to system dynamics modeling with dynamo. Pegasus Communications
19. Sterman J (2000) Business dynamics. McGraw-Hill, Irwin
20. Sterman J (ed) (2002) Special Issue of the Syst Dyn Rev
21. Sterman J, Sweeney L (2007) Understanding public complacency about climate change. *Clim Chang* 80(3–4):213–238
22. Warren K (2002) Competitive strategy dynamics. Wiley, Chichester
23. Webster M et al (2003) Uncertainty analysis of climate change and policy response. *Climat Chang* 61:295–320

### Books and Review

Houghton J (2004) Global warming: The complete briefing, 3rd edn. Cambridge University Press, Cambridge

## System Dynamics Models, Optimization of

BRIAN DANGERFIELD

Centre for OR & Applied Statistics, Salford Business School, University of Salford, Salford, UK

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Optimization as Calibration](#)

[Optimization of Performance \(Policy Optimization\)](#)

[Examples of SD Optimization Reported in the Literature](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Econometrics** A statistical approach to economic modeling in which all the parameters in the structural equations are estimated according to a ‘best fit’ to historical data.

**Maximum likelihood** A statistical concept which underpins calibration optimization and which generates the most likely parameter values; it is equivalent to the parameter set which minimizes the chi-square value.

**Objective function** See **Payoff** below.

**Optimization** The process of improving a model’s results in terms of either an aspect of its performance or by calibrating it to fit reported time series data.

**Payoff** A formula which expresses the objective, say, maximization of profits, minimization of costs or minimization of the differences between a model variable and historical data on that variable.

**Zero-one parameter** A parameter which is used as a multiplier in a policy equation and serves the effect of bringing in or removing a particular influence in determining the optimal policy.

### Definition of the Subject

The term ‘optimization’ when related to system dynamics (SD) models has a special significance. It relates to the mechanism used to improve the model vis-à-vis a criterion. This collapses into two fundamentally different intentions. Firstly one may wish to improve the model in terms of its performance. For instance, it may be desired to minimize overall costs of inventory whilst still offering a satisfactory level of service to the downstream customer. So the criterion here is cost, and this would be minimized after searching the parameter space related to service level. The direction of need may be reversed and maximization may be desired as, for instance, if one had a model of a firm and wished to maximize profit subject to an acceptable level of payroll and advertising costs. Here the parameter space being explored would involve both payroll and advertising parameters. This type of optimization might be described generically as *policy optimization*.

Optimization of performance is also the *raison d’être* of other management science tools, most notably mathematical programming. But such tools are usually static: they offer the ‘optimum’ resource allocation given a set of constraints and a performance function to either maximize or minimize. These models normally relate to a single time point and may then need to be re-run on a weekly or monthly basis to determine a new optimal resource allocation. In addition, these models are often linear (certainly so in the case of linear programming), whereas SD models are usually non-linear. So the essential differences are that SD model optimization for performance involves both a dynamic and a non-linear model.

A separate improvement to the model may be sought where it is required to fit the model to past time series data. Optimization here involves minimizing a statistical function which expresses how well the model fits a time-series of data pertaining to an important model variable. In other words a vector of parameters are explored with a view to determining the particular parameter combination which



offers the best fit between the chosen important model variable and a past time series data set of this variable. This type of optimization might be generically termed *model calibration*. If *all* the parameters in the SD model are determined in this fashion then the process is equivalent to the technique of econometric modeling. A good comparison between system dynamics and econometric modeling can be found in Meadows and Robinson [12].

### Optimization as Calibration

In these circumstances we wish to determine optimal parameters, those which, following a search of the parameter space, offer the best fit of a particular model variable to a time series dataset on that variable taken from real world reporting.

As an example consider a variation of the one of the epidemic models which are made available with the Vensim™ software. The stock-flow diagram is presented as Fig. 1.

In this epidemiological system members of a susceptible population become infected and join the infected population. Epidemiologists call this an S–I model. It is a simpler variation of the S–I–R model which includes recovered (R) individuals.

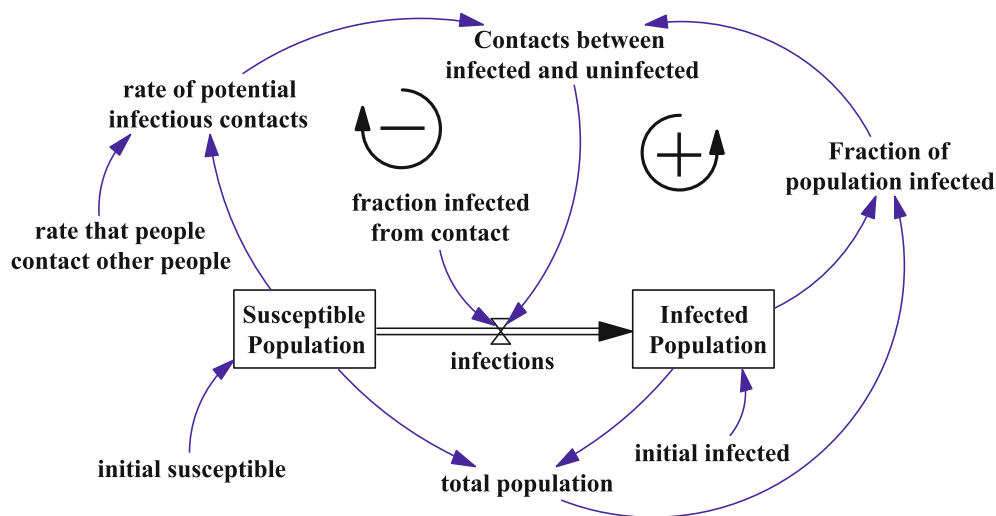
Suppose some data on new infections (at intervals of five days) are available covering 25 days of a real-world epidemic. The model is set with a time horizon of 50 days which is consistent with, say, a flu epidemic or an infectious outbreak of dysentery in a closed population such

as a cruise ship. The ‘current’ run of the model is shown in Fig. 2, with the real-world data included for comparison.

Clearly there is not a very good correspondence between the actual data and the model variable for the infection rate (infections). We wish to achieve a better calibration, and so there is a need to select relevant parameters through which the calibration optimization can be performed over. Referring back to Fig. 1, we can see that the *fraction infected from contact* and the *rate that people contact other people* are two possible parameters to consider. The initial infected and initial susceptible are also parameters of the model in the strict sense of the term, but we will ignore them on this occasion. In this model the *initial infected* is 10 persons and *initial susceptibles* number 750,000 persons.

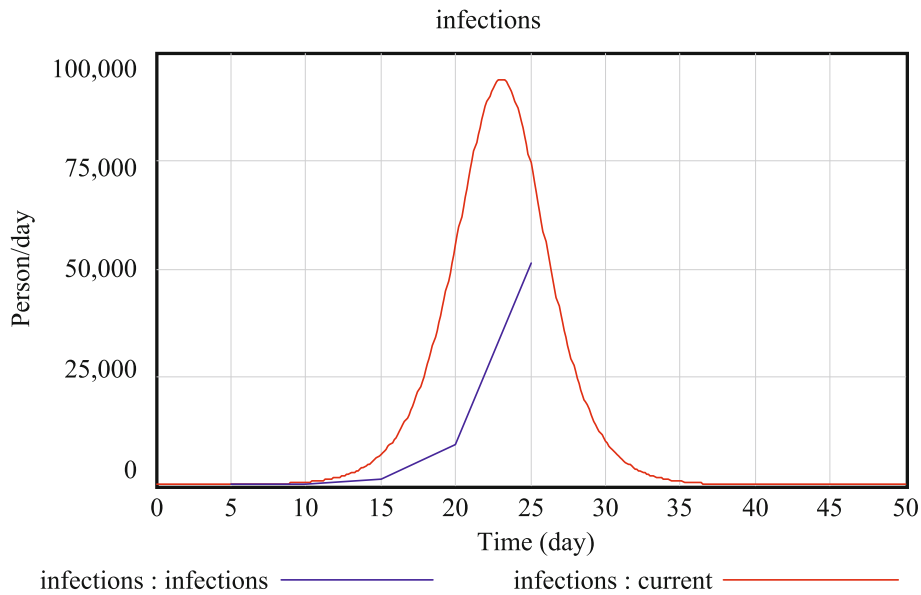
The chosen value for the *fraction infected from contact* is 0.1, while that for the *rate that people contact other people* is 5.0. The former is a dimensionless number while the latter is measured as a fraction per day (1/day). This is obtained from consideration of the *rate of potential infectious contacts* (persons/day) as a proportion of the *susceptible population* (persons).

The optimization process for calibration involves reading into the model the time series data, in this case on new infections, and, secondly, determining the range for the search in parameter space. There is usually some basic background knowledge which allows a sensible range to be entered. For instance, a probability can only be specified between 0 and 1.0. In this case we have chosen to specify



System Dynamics Models, Optimization of, Figure 1  
Stock-flow diagram for a simple epidemic model





System Dynamics Models, Optimization of, Figure 2  
Current (base) run of the model and reported data on infections

the ranges as follows:

$$0.03 \leq \text{fraction infected from contact} \leq 0.7$$
$$2 \leq \text{rate that people contact other people} \leq 10 .$$

A word of warning is necessary in respect of optimizing delay parameters. Because there is a risk of mathematical instability in the model if the value of DT (the TIME STEP) is too large relative to the smallest first-order delay constant, it is important to ensure the TIME STEP employed in the model is sufficiently small to cope with delay constant values which may be reached during the search of the delay parameter space. In other words ensure the minimum number for the search range on the delay parameter is at least double the value of the TIME STEP.

**Maximum Likelihood Estimation and the Payoff Function**

The optimization process involves a determination of what are termed statistically as maximum likelihood estimates. In Vensim™ this is achieved by maximizing a payoff function. Initially this is negative and the optimization process should ensure this becomes less negative. An ideal payoff value, after optimization, would be zero. A weighting is needed in the payoff function too, but for calibration optimization this is normally 1.0. Driving the payoff value to be larger by making it less negative has parallels with the operation with the simplex algorithm common in linear programming. This algorithm was conceived initially

for problems where the objective function was to be minimized. Its use on maximization problems is achieved by minimizing the negative of the objective function.

During the calibration search, Vensim™ takes the difference between the model variable and the data value, multiplies it by the weight, squares it and adds it to the error sum. This error sum is minimized. Usually data points will not exist at every time point in the model. Here the model TIME STEP is 0.125 (1/8th), but let us assume that reported data on new infections have been made available only at times  $t = 5, 10, 15, 20$  and  $25$  so the sum of squares operation is performed only at these five time points.

The data are shown as Table 1.

System Dynamics Models, Optimization of, Table 1  
Data used for calibration experiment

| Time       | 5  | 10  | 15   | 20   | 25     |
|------------|----|-----|------|------|--------|
| Infections | 30 | 230 | 1400 | 9500 | 51 400 |

**The Recording Point for Reported Data**

System dynamics models differentiate between stock and flow variables and the software used for simulating such models advances by a small constant TIME STEP (also known as DT). This has implications for the task of fitting real-world reported data to each type of system dynamics model variable. The following is the issue: at what point

in a continuum of time steps should the reported data be recorded at? This is important because the reported data has to be read into the model to be compared with the simulated data. The answer will be different for stock and flow variables.

Where the reported data relate to a stock variable the appropriate time point for recording will be known. If it is recorded at the end of the day (say a closing bank balance) then the appropriate point for data entry in the model will be the beginning of the next day. Thus the first data point above is at time  $t = 5$  (5.00) and would, if it were a stock, correspond to a record taken at the very end of time period 4.

However, if the data relate to a flow variable, as in the case of new infections here, the number is the total new infections which have occurred over the entire time unit (day, week, month etc.) and so there is a decision to be reached as to which time point the data are entered at. This is because the TIME STEP (DT) is hardly ever as large as the basic time unit which the model is calibrated in. The use of 5 (10, 15 etc.) above implies that the data on new infections over the period of time  $t = 0$  to  $t = 5$  are compared with the corresponding model variable at time  $5 + 1 * DT$  (and the new infections over the period  $t = 5$  to  $t = 10$  at time  $10 + 1 * DT$  etc.). A more appropriate selection might be towards the end of the 5-day time period. Following the example above using a TIME STEP = 0.125, this might be at time  $4 + 7 * DT$  (that is at 4.875).

### Calibration Optimization Results

Based upon the data on new infections shown above and the chosen ranges for the parameter search, the following output is obtained (Table 2). After 114 simulations the optimized values for our two parameters are shown to be

**System Dynamics Models, Optimization of, Table 2**  
Results from the calibration optimization

|                                                                                                                                                                                           |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Initial point of search<br>fraction infected from contact = 0.1<br>rate that people contact other people = 5<br>Simulations = 1<br>Pass = 0<br>Payoff = $-2.67655e + 009$                 |
| Maximum payoff found at:<br>fraction infected from contact = 0.0794332<br>*rate that people contact other people = 5.11568<br>Simulations = 114<br>Pass = 6<br>Payoff = $-1.06161e + 006$ |

0.08 and 5.12 and the payoff is over 2500 times larger (less negative). Replacing the original parameters with the optimized values reveals the result shown in Fig. 3. To take things further we may wish to put confidence intervals on the estimated parameters. One way of accomplishing this is by profiling the likelihood and is described in Dangerfield and Roberts [3].

### Avoid Cumulated Data

There might be a temptation to optimize parameters against cumulated data when the data are reported essentially as a flow, as is the case here. Were the data to be cumulated we would obtain as shown in Table 3.

The results from this optimization are shown in Table 4. The ranges for the parameter space search are kept the same but the payoff function now involves a comparison of the model variable *infected population* with the corresponding cumulated data. Figure 4 shows the resultant fit to infected population is good, but that is manifestly not borne out when we consider the plot of infections obtained from the same optimization run (Fig. 5).

The reason for this is rooted in statistics. The maximum likelihood estimator is equivalent to the chi-squared statistic. This in turn assumes that each expected data value is independent. A cumulated data series would not exhibit this property of independence.

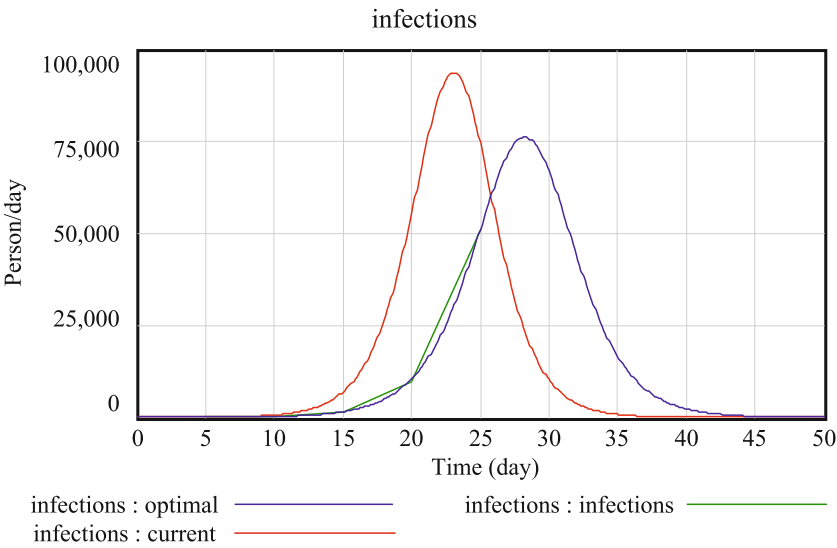
As an aside it is worth pointing out that this model, with suitable changes to the variable names and the time constants involved, could equally represent the diffusion

**System Dynamics Models, Optimization of, Table 3**  
Cumulated reported data for the infected population

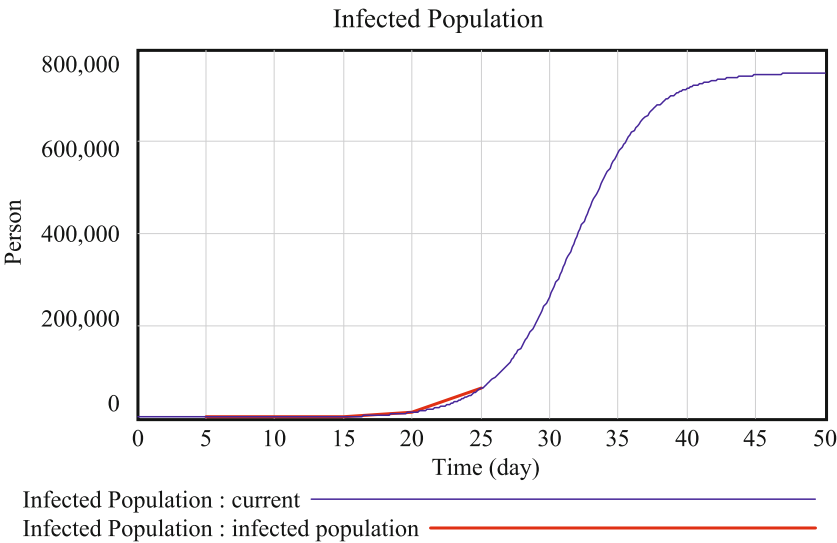
| Time                | 5  | 10  | 15   | 20     | 25     |
|---------------------|----|-----|------|--------|--------|
| Infected population | 30 | 260 | 1660 | 11 160 | 62 560 |

**System Dynamics Models, Optimization of, Table 4**  
Results from the calibration using cumulated data

|                                                                                                                                                                                                                      |
|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Initial point of search<br>fraction infected from contact = 0.1<br>rate that people contact other people = 5<br>Simulations = 1<br>Pass = 0<br>Payoff = $-2.48206e + 011$                                            |
| Maximum payoff found at:<br>fraction infected from contact = 0.0726811<br>*rate that people contact other people = 4.96546<br>Simulations = 145<br>Pass = 6<br>Payoff = $-212 645$<br>The final payoff is $-212 645$ |



System Dynamics Models, Optimization of, Figure 3  
Reported data on infections and optimized (calibrated) model; the base case (current) is reproduced for reference



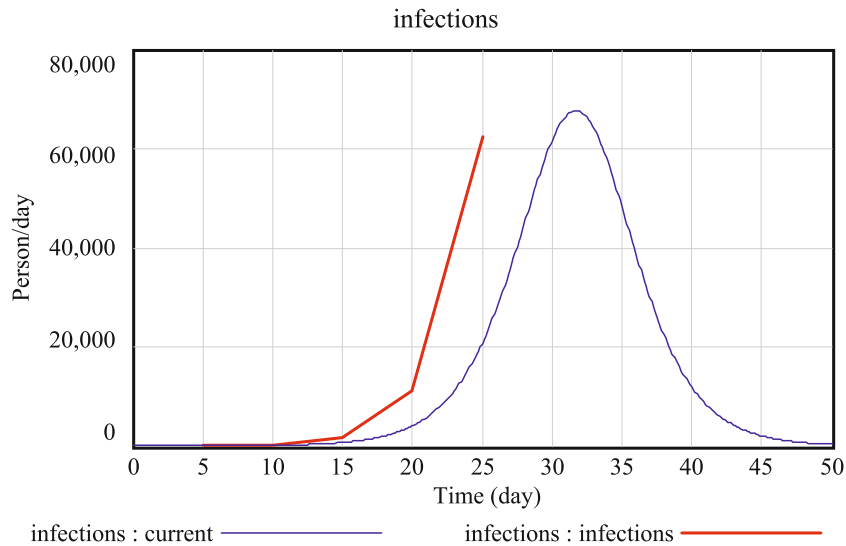
System Dynamics Models, Optimization of, Figure 4  
The cumulative model variable (infected population) together with reported data

of a new product into a virgin market. In systems terms the structures are equivalent. The *fraction infected from contact* is the same as, say, the fraction reached by word of mouth or advertising and the *rate that people contact other people* is a measure of the potential interactions at which new products might be mentioned amongst the members of the relevant market segment. An *infected population* is equivalent to a customer base, the number of adopters of the relevant product. So it is possible to shed light

on important real-world marketing parameters through a calibration optimization of models of this general structure.

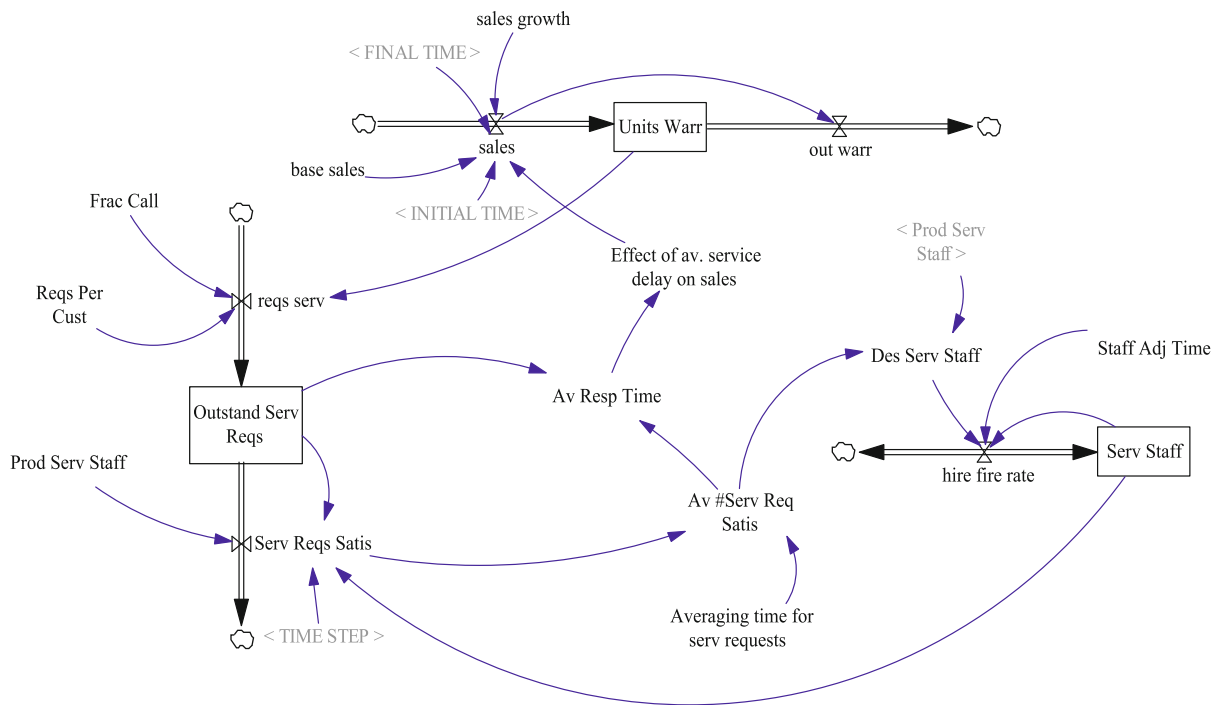
Optimization of Performance (Policy Optimization)

An example model is to be used to illustrate the process of optimization to improve the performance of the system, and this is illustrated in Fig. 6. It concerns the service re-



System Dynamics Models, Optimization of, Figure 5

The corresponding fit to infections is poor



System Dynamics Models, Optimization of, Figure 6

Model of service delays for durable goods under warranty

quirements which can arise following the sale of a durable good. These items are typically sold with a 12-month warranty and during this time the vendor is obliged to offer service if a customer calls for it. In this particular case the vendor is not being responsive in terms of staffing the ser-

vice section. The result is that as sales grow the increasing number of service requests is putting pressure on the service personnel. The delay in responding to service calls also increases and the effect of this is that future sales are depressed because of the vendor's acquired reputation for

poor service response. The basic behavior mode is overshoot and collapse.

In the model depicted in Fig. 6, the growth process is achieved by a RAMP function which causes sales of the good to increase linearly by 20 units per month from a base of 500 units per month.

The payoff function is restricted to the variable *Sales*. However, this need not be the case. Where a number of variables might be options in a payoff function, it is possible to assign weights to each such that the sum of the weights is 1.0 (or 100). The optimization process will then proceed with the software accumulating a weighted payoff which it will attempt to maximize. Weights are positive when more is better and negative when less is better.

Policy Experiment No. 1

Here it is decided to try to improve the productivity of the service staff. Currently they manage, on average, to respond to 120 calls per operative per month. It may be an option to improve their productivity by, say, providing them with hand-held devices which direct each operative from one call to the next – calls which may have arisen since setting out from their base. In this way their call routing is improved.

The optimization parameter is *Prod Serv Staff*, and we select an upper limit for the search range of 240 calls per person per month. The chosen performance variable is *Sales*, since we wish to maximize this – or at least not have it overly depressed by poor response times. The results are shown in Table 5. We see that the payoff is increased and that the optimum productivity is a modest increase of 2.6 requests per month, on average. This should be easily achievable and perhaps without expenditure on high-tech devices. The graphical output for sales is shown in Fig. 7.

For comparison, the effect of increasing the productivity to as high as 150 calls per month, on average, is

also shown. This would represent an increase of 25% and would be much more difficult to accomplish. Here the benefit of optimization is highlighted. A modest increase in productivity returns a visibly improved sales performance (although the basic behavior mode is unchanged), whilst a much greater productivity increase offers little extra benefit for the effort and cost involved in improving productivity.

Policy Experiment No. 2

Another approach to policy optimization involves the use of a zero-one parameter which has the effect of either including or excluding an influence on policy. Suppose it was thought that the quantity of product units in warranty should exert an influence on the numbers of service personnel hired (or fired). The equation for the desired number of service staff (*Des Serv Staff*) can be expressed as:

$$\text{Des Serv Staff} = \text{"Av \#Serv Req Satis"}/\text{Prod Serv Staff} * \text{trigger} + (\text{"Av \#Serv Req Satis"}/\text{Prod Serv Staff}) * (\text{Units Warr}/\text{initial units in warranty}) * (1 - \text{trigger}).$$
 (Units: Persons).

The *trigger* variable is initially set to 1.0 and so the more sophisticated policy is not active. The optimization run results are shown in Table 6. Clearly there is benefit from including the more sophisticated policy which takes into account the current numbers of product units in warranty.

The graphical output is unequivocal (Fig. 8). Sales are continuously increasing when the recruitment policy for service personnel takes into account the number of product units in warranty. The depressive effect on sales of poor service performance is non-existent.

Whilst this might seem an obvious policy, it is surprising how easily the naïve alternative might be accepted without question. The number of calls a typical operative can manage each month is well known along with the (his-

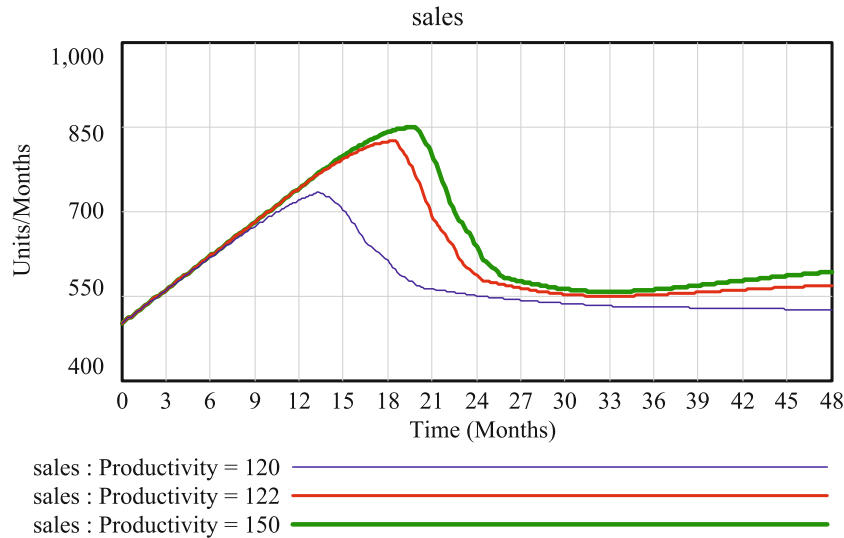
System Dynamics Models, Optimization of, Table 5  
Optimization results for the productivity of the service staff

|                                                                                                           |
|-----------------------------------------------------------------------------------------------------------|
| Initial point of search<br>Prod Serv Staff = 120<br>Simulations = 1<br>Pass = 0<br>Payoff = 27 743.5      |
| Maximum payoff found at:<br>*Prod Serv Staff = 122.647<br>Simulations = 27<br>Pass = 3<br>Payoff = 29 915 |

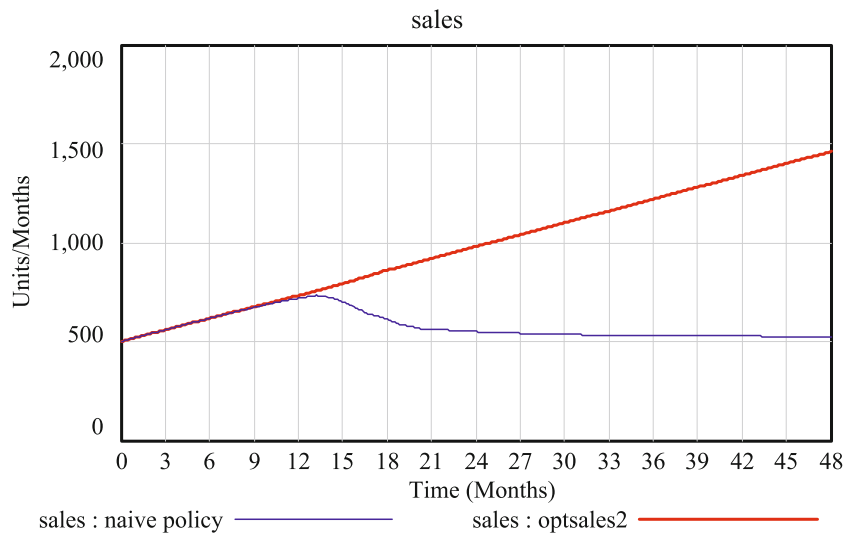
System Dynamics Models, Optimization of, Table 6  
Optimization results from selection of policy drivers

|                                                                                               |
|-----------------------------------------------------------------------------------------------|
| Initial point of search<br>trigger = 1<br>Simulations = 1<br>Pass = 0<br>Payoff = 27 743.5    |
| Maximum payoff found at:<br>*trigger = 0<br>Simulations = 13<br>Pass = 3<br>Payoff = 47 090.9 |





**System Dynamics Models, Optimization of, Figure 7**  
Plots of sales achieved for differing productivities



**System Dynamics Models, Optimization of, Figure 8**  
Comparison of sales from two different policy drivers

torical) number of service requests satisfied. Hence, the desired number of staff is more or less fixed. This comes undone when there is a growth in the number of products sold. In this different environment such a simplistic policy can, as shown, lead to overshoot and collapse. Notice needs to be taken of the changing number of product units in warranty in order that a more effective system performance is achieved.

The above experiments are illustrative only, and there is no intention of over-working a simple teaching model in order to uncover an ideal policy. In the case of pol-

icy optimization a wide range of possible alternatives exists. Indeed, a process of learning naturally arises through carrying out repeated optimization experiments with the model [1].

### Examples of SD Optimization Reported in the Literature

Amongst the earliest work in this area the writings of Kelo-harju are worthy of mention. He contributed a number of papers on the topic in the pages of *Dynamica*. See, for

example, Keloharju [9]. His work brought the concept to prominence but he did not employ the method on anything other than problems described in text books or postulated by himself. For instance, an application of optimization to the project model contained in Richardson and Pugh's [13] text is contained in Keloharju and Wolstenholme [11]. A statement of the method together with some textbook examples is also available [10]. Additionally, an overview of the methods and their deployment on textbook examples has been contributed by the current author [3]. Finally, there is an example of optimization applied to defence analysis. Again though it is a standard defence model – the armored advance model – rather than any real-world study [14].

Retaining the emphasis on textbook problems for the moment, Duggan [6] employs Coyle's model [1] of the *Domestic Manufacturing Company* to illustrate the methods of multi-objective optimization – an advance over standard SD optimization with its single objective function. The concept of multiple objectives arises from multi-criteria decision-making where a situation can be judged on more than one performance metric. While a multi-objective payoff function can be formulated using a set of weights, it is argued that the selection of the weights is very individual-specific. The multi-objective approach – underpinned by the methods of genetic algorithms – rests upon determining a Pareto-optimal situation, defined as one where no improvement is possible without making some other aspect worse. In other words the method strives for an optimal solution which is not dominated by any other solution. The author demonstrates the approach combining two objectives in the model: one for the differences between desired stock and actual and another between desired backlog and actual.

In terms of applications to real-world problems, the current author has also used the methods of optimization in research conducted in connection with modeling the epidemiology of HIV/AIDS. Fitting a model of AIDS spread to data was carried out for a number of European countries [2,4]. The optimized parameters furnished support for some of the features of AIDS epidemiology which, at the time, were being uncovered by other branches of science. For example, the optimized output revealed that a U-shaped profile of infectiousness in a host was necessary in order to achieve a best fit to data on new AIDS cases. This infectiousness profile was also evidenced by virologists who had analyzed patients' blood and other secretions on a longitudinal basis.

Within this strand of research, a much more complex optimization was performed using American data on transfusion-associated AIDS cases [5]. The purpose here

was to estimate the parameters for a number of plausible statistical HIV incubation distributions. Given the nature of the data, the point of infection could be quite accurately determined, but two difficulties were evident: the data were right-censored and the number receiving infected transfusions in each quarter was unknown. However, the SD optimization could estimate this number as part of the process, in addition to estimating parameters of the incubation distribution. The best fit was found to be a three stage distribution similar to the gamma and one which accorded with the high-low-high U-shaped infectiousness profile which was receiving support from a number of sources.

In the marketing domain Graham and Ariza [8] carried out an optimization on a system dynamics model which was designed to shed light on the allocations to make from a marketing budget in a high-tech client firm. Assuming the budget was fixed, the task was to optimize the allocations across more than 90 'buckets' – combinations of product lines, marketing channels and types of marketing. However, these were not discrete: advertising on one product line might have crossover effects on another and the impacts could propagate over a period of time. One major conclusion for this firm was that the advertising allocation should be increased markedly. In general intuitive allocations were shown to fall short of the ideal: they were directionally correct but magnitudes fell short often by factors of three or four.

## Future Directions

A primary aim must be to see more published work which describes optimization studies carried out on real-world SD applications. There may be frequent use of optimization in consulting assignments but such activities are rarely published. The references herein suggest that, thus far, outside of unpublished work, the number may be three at most. Whilst software requirements may have inhibited use of SD optimization in the past, there are now no computational barriers to its use and it is to be hoped that in future this quite powerful analytical tool in SD will feature in more application studies.

An advance in the methodology itself has been developed by Duggan [7] and this is a promising pointer for the future. Based on genetic algorithms, it is best suited to the class of SD problems that are agent-based and this highlights a slight limitation. Traditional optimization takes the policy equations as given and explores the parameter space to determine an optimal policy. Instead he has offered an approach which searches over both parameter space and policy (strategy). Theoretically there is no limit

to the number of strategies which can be evaluated in this approach, although the user has to define a set in advance of the runs. Under a conventional optimization approach a limited tilt at this is possible using the zero-one parameter method suggested above, although this would restrict the enumerated strategies to two only. Duggan demonstrates the new approach using a classic SD problem: the four agent beer-game. We await its use in a real-world application.

## Bibliography

1. Coyle RG (1996) System dynamics modelling: a practical approach. Chapman & Hall, London
2. Dangerfield BC, Roberts CA (1994) Fitting a model of the spread of AIDS to data from five European countries. In: O.R. Work in HIV/AIDS, 2nd edn. Operational Research Society, Birmingham, pp 7–13
3. Dangerfield BC, Roberts CA (1996) An overview of strategy and tactics in system dynamics optimisation. *J Oper Res Soc* 47(3):405–423
4. Dangerfield BC, Roberts CA (1996) Relating a transmission model of AIDS spread to data: some international comparisons. In: Isham V, Medley G (eds) Models for infectious human diseases: Their structure and relation to data. Cambridge University Press, Cambridge, pp 473–476
5. Dangerfield BC, Roberts CA (1999) Optimisation as a statistical estimation tool: an example in estimating the AIDS treatment-free incubation period distribution. *Syst Dyn Rev* 15(3):273–291
6. Duggan J (2005) Using multiple objective optimisation to generate policy insights for system dynamics models. In: Proceedings of the international system dynamics conference, Boston. System Dynamics Society. (CD-ROM)
7. Duggan J (2008) Equation-based policy optimisation for agent-oriented system dynamics models. *Syst Dyn Rev* 24(1):97–118
8. Graham AK, Ariza CA (2003) Dynamic, hard and strategic questions: using optimisation to answer a marketing resource allocation question. *Syst Dyn Rev* 19(1):27–46
9. Keloharju R (1977) Multi-objective decision models in system dynamics. *Dynamica* 3(1):3–13 and 3(2):45–55
10. Keloharju R, Wolstenholme EF (1988) The basic concepts of system dynamics optimisation. *Syst Pract* 1:65–86
11. Keloharju R, Wolstenholme EF (1989) A case study in system dynamics optimisation. *J Oper Res Soc* 40(3):221–230
12. Meadows DM, Robinson JM (1985) The electronic oracle. Wiley, Chichester (Now available from the System Dynamics Society, Albany NY)
13. Richardson GP, Pugh AL (1981) An introduction to system dynamics modelling with DYNAMO. MIT Press, Cambridge (Now available from Pegasus Communications, Waltham, MA)
14. Wolstenholme EF, Al-Alusi AS (1987) System dynamics and heuristic optimisation in defence analysis, *Syst Dyn Rev* 3(2):102–115

## System Dynamics and Organizational Learning

KAMBIZ MAANI

Chair in Systems Thinking and Practice,  
The University of Queensland, Brisbane, Australia

### Article Outline

Glossary

Definition of the Subject

Introduction

Systems Thinking and Modeling Methodology

Organizational Learning

Modeling for Organizational Learning

Future Directions

Bibliography

### Glossary

**Stock** In system dynamics stock is a concept representing accumulation and the state of a variable, such as, assets, inventory, capacity, reputation, morale etc. Stock can be measured at any point of time. In mathematical terms, stock is the sum over time (integral) of one or more flows.

**Flow** Flow or rate represents change or movement in a stock such as, buying assets, building inventories, adding capacity, losing reputation or morale, etc. Flow is measured as “per unit of time” like hiring rate (employees hired per year, production rate (units made per day), or rainfall (inches of rain per month).

**Causal loops** Causal loops (model) are visual maps that connect a group of variables with known or hypothesized cause and effect relationships. A causal loop can be open or closed. Causal loops can be used for complex problem solving/decision making, consensus building, conflict resolution, priority setting and group learning.

**Feedback** In a cause and effect chain (system), feedback is a signal from the effect/s to cause/s as to its/their influence on downstream effect/s. Feedback can be information, decision or action. For example, if *X* causes or changes *Y*, *Y* in turn could influence or change *X* directly or through other intervening variables. This creates a *closed* “causal loop” with either a positive or amplifying feedback (Reinforcing – *R*) or a negative feedback with damping, counteracting or (Balancing – *B*) effect.

**Delay** Cause and effect relationships are often not close in time or space. The lapse time between a cause and its effect is called a systems delay or simply delay. Because

some delays in physical, natural and social systems are rather long they mask the underlying or earlier causes when effects become evident. This provides confusion and unintended consequences, especially in social systems, such as economics, education, immigration, judicial systems, etc.

**Reference Mode** Reference mode is the actual/observed pattern of a key variable of interest to decision makers or policy analysts. It represents the actual behavior of a variable over time which is used to compare with the simulated pattern of the same variable generated by a simulation model to validate the accuracy of the model.

**Simulation** A computer tool and methodology for modeling complex situations and challenging problems where mathematical tools fail to operate.

**Microworld** Microworlds are simulation models of real systems such as a firm, a hospital, a market, or a production system. They provide a “virtual” world where decision makers can test and experiment their policies and strategies in a laboratory environment before implementation. Microworlds are constructed using system dynamic software with user friendly interfaces.

**Leverage** Leverage refers to decisions and actions for change and intervention which have the highest likelihood of lasting and sustainable outcomes. Leverage decisions are best reached by open discussion after the group develops a deep understanding of system dynamics through a causal loop or stock & flow modeling process.

**Systems thinking** Systems thinking is a paradigm for viewing reality based on the primacy of the whole and relationships. It is one of the key capabilities (disciplines) for organizational learning [30]. Systems Thinking consists of a series of conceptual and modeling tools such as behavior over time, causal loop diagrams and systems archetypes. These tools reveal cause and effect dynamics over time and assist understanding of complex, non-linear, and counter-intuitive behaviors in all systems – physical, natural and social.

## Definition of the Subject

System dynamics (SD) is “a methodology for studying and managing complex feedback systems... While the word system has been applied to all sorts of situations, feedback is the differentiating descriptor here. Feedback refers to the situation of  $X$  affecting  $Y$  and  $Y$  in turn affecting  $X$  perhaps through a chain of causes and effects... Only the study of the whole system as a feedback system will lead to correct results.” [36]

Sterman ([35], p 4) defines System Dynamics as “a method to enhance learning in complex systems”. “System dynamics is fundamentally interdisciplinary... It is grounded in the theory of nonlinear dynamics and feedback control developed in mathematics, physics, and engineering. Because we apply these tools to the behavior of human as well as physical and technical systems, system dynamics draws on cognitive and social psychology, economics, and other social sciences.”

Wolstenholme’s [40] offers the following description for system dynamics and its scope:

A rigorous way to help thinking, visualizing, sharing, and communication of the future evolution of complex organizations and issues over time; for the purpose of solving problems and creating more robust designs, which minimize the likelihood of unpleasant surprises and unintended consequences; by creating operational maps and simulation models which externalize mental models and capture the interrelationships of physical and behavioral processes, organizational boundaries, policies, information feedback and time delays; and by using these architectures to test the holistic outcomes of alternative plans and ideas; within a framework which respects and fosters the needs and values of awareness, openness, responsibility and equality of individuals and teams.

## Organizational Learning

Organizational learning is the ability of organizations to enhance their collective capacity to learn and to act, harmoniously. According to Senge [30] “Real learning gets to the heart of what it means to be human. Through learning we re-create ourselves. Through learning we become able to do something we never were able to do. Through learning we re-perceive the world and our relationship to it. Through learning we extend our capacity to create, to be part of the generative process of life. There is within each of us a deep hunger for this type of learning.” Organizational learning extends this learning to the organization and its members.

## Introduction

### History of System Dynamics

(This section is due to US Department of Energy website) “System dynamics was created during the mid-1950s by Professor Jay W. Forrester of the Massachusetts Institute of Technology. Forrester arrived at MIT in 1939 for graduate study in electrical engineering. His first research assistantship put him under the tutelage of Professor Gordon

Brown, the founder of MIT's Servomechanism Laboratory. Members of the MIT Servomechanism Laboratory, at the time, conducted pioneering research in feedback control mechanisms for military equipment. Forrester's work for the Laboratory included traveling to the Pacific Theatre during World War II to repair a hydraulically controlled radar system installed aboard the aircraft carrier Lexington. The Lexington was torpedoed while Forrester was on board, but not sunk.

At the end of World War II, Jay Forrester turned his attention to the creation of an aircraft flight simulator for the US Navy. The design of the simulator was cast around the idea, untested at the time, of a digital computer. As the brainstorming surrounding the digital aircraft simulator proceeded, however, it became apparent that a better application of the emerging technology was the testing of computerized combat information systems. In 1947, the MIT Digital Computer Laboratory was founded and placed under the direction of Jay Forrester. The Laboratory's first task was the creation of WHIRLWIND I, MIT's first general-purpose digital computer, and an environment for testing whether digital computers could be effectively used for the control of combat information systems. As part of the WHIRLWIND I project, Forrester invented and patented coincident-current random-access magnetic computer memory. This became the industry standard for computer memory for approximately twenty years. The WHIRLWIND I project also motivated Forrester to create the technology that first facilitated the practical digital control of machine tools.

After the WHIRLWIND I project, Forrester agreed to lead a division of MIT's Lincoln Laboratory in its efforts to create computers for the North American SAGE (Semi-Automatic Ground Environment) air defense system. The computers created by Forrester's team during the SAGE project were installed in the late 1950s, remained in service for approximately twenty-five years, and had a remarkable "up time" of 99.8%.

Forrester's seminal book *Industrial Dynamics* [11] "is still a significant statement of philosophy and methodology in the field. Since its publication, the span of applications has grown extensively and now encompasses work in

- corporate planning and policy design
- public management and policy
- biological and medical modeling
- energy and the environment
- theory development in the natural and social sciences
- dynamic decision making
- complex nonlinear dynamics" [36]

## Systems Thinking and Modeling Methodology

System Dynamics is one of the five phases of systems thinking and modeling intervention methodology [6,21]. These distinct but related phases are as follows:

1. Problem structuring;
2. Causal loop modeling;
3. System dynamics modeling;
4. Scenario planning and modeling;
5. Implementation and organizational learning (learning lab).

These phases follow a process, each involving a number of steps, as outlined in Table 1. This process does not require all phases to be undertaken, nor does each phase require all the steps listed. Which phases and steps are included in a particular project or intervention depends on the issues or problems that have generated the systems enquiry and the degree of effort that the organization is prepared to commit to.

## System Dynamics Modeling

This phase follows the causal modeling phase. Although it is possible to go into this phase directly after problem structuring, performing the causal modeling phase first will enhance the conceptual rigor and learning power of the systems approach. The completeness and wider insights of systems thinking is generally absent from other simulation modeling approaches, where causal modeling does not play a part. The following steps are generally followed in the system dynamics modeling phase.

1. Develop a high-level map or systems diagram showing the main sectors of a potential simulation model, or a 'rich picture' of the main variables and issues involved in the system of interest.
2. Define variable types (e.g. stocks, flows, converters, etc.) and construct stock flow diagrams for different sectors of the model.
3. Collect detailed, relevant data including media reports, historical and statistical records, policy documents, previous studies, and stakeholder interviews.
4. Construct a computer simulation model based on the causal loop diagrams or stock-flow diagrams. Identify the initial values for the stocks (levels), parameter values for the relationships, and the structural relationships between the variables using constants, graphical relationships and mathematical functions where appropriate. This stage involves using specialized computer packages like STELLA, *ithink*, VENSIM, POWERSIM, DYSMAP, COSMIC and Consideo.



System Dynamics and Organizational Learning, Table 1

The five phase process of systems thinking and modeling (Source: [6])

| Phases | Steps                                      |                                                                                                                                                                                                                                                                                                                                                                                                      |
|--------|--------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1      | Problem structuring                        | Identify problems or issues of concern to management,<br>Collect preliminary information and data                                                                                                                                                                                                                                                                                                    |
| 2      | Causal loop modeling                       | Identify main variables,<br>Prepare behavior over time graphs (reference mode),<br>Develop causal loop diagram (influence diagram),<br>Analyze loop behavior over time and identify loop types,<br>Identify system archetypes,<br>Identify key leverage points,<br>Develop intervention strategies                                                                                                   |
| 3      | <b>System dynamic modeling</b>             | Develop a systems map or rich picture,<br>Define variable types and construct stock-flow diagrams,<br>Collect detailed information and data,<br>Develop a simulation model,<br>Simulate steady -state/stability conditions,<br>Reproduce reference mode behavior (base case),<br>Validate the model,<br>Perform sensitivity analysis,<br>Design and analyze policies,<br>Develop and test strategies |
| 4      | Scenario planning and modeling             | Plan general scope of scenarios,<br>Identify key drivers of change and keynote uncertainties,<br>Construct forced and learning scenarios,<br>Simulate scenarios with the model,<br>Evaluate robustness of the policies and strategies                                                                                                                                                                |
| 5      | Implementation and organizational learning | Prepare a report and presentation to management team,<br>Communicate results and insights of proposed intervention to stakeholders,<br>Develop a microworld and learning lab based on the simulation model,<br>Use learning lab to examine mental models and facilitate                                                                                                                              |

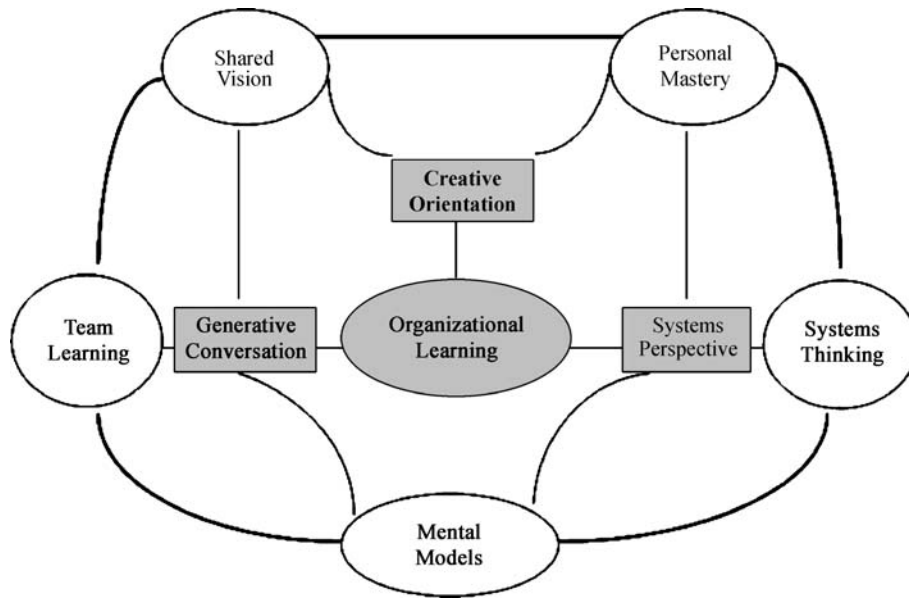
5. Simulate the model over time. Select the initial value for the beginning of the simulation run, specify the unit of time for the simulation (e. g. hour, day, week, month, year, etc.). Select the simulation interval (DT) (e. g. 0.25, 0.5, 1.0) and the time horizon for the simulation run (i. e. the length of the simulation). Simulate model stability by generating steady state conditions.
6. Produce graphical and tabular output for the base case of the model. This can be produced using any of the computer packages mentioned above. Compare model behavior with historical trends or hypothesized reference modes (behavior over time charts).
7. Verify model equations, parameters and boundaries, and validate the model's behavior over time. Carefully inspect the graphical and tabular output generated by the model.
8. Perform sensitivity tests to gauge the sensitivity of model parameters and initial values. Identify areas of greatest improvement (key leverage points) in the system.
9. Design and test policies with the model to address the issues of concern to management and to look for system improvement.
10. Develop and test strategies (i. e. combinations of functional policies, for example operations, marketing, finance, human resources, etc.).

### Organizational Learning

(This section is adapted from [21])

Peter Senge, who popularized the concept through his seminal book: *The Fifth Discipline* [30], describes a learning organization as one 'which is continually expanding its ability to create its future'. He identifies five core capabilities (disciplines) of the learning organization that are derived from three "higher orientations": creative orientation; generative conversation; and systems perspective. "The reality each of us sees and understands depend on what we believe is there. By learning the principles of the five disciplines, teams begin to understand how they can think and inquire that reality, so that they can collaborate in discussions and in working together create the results that matter [to them]."

As Fig. 1 shows the learning organization capabilities are dynamically interrelated, and collectively they lead to organizational learning.



System Dynamics and Organizational Learning, Figure 1  
The core capabilities of a learning organization (Source: [19])

Senge maintains that *Creative orientation* is the source of a genuine desire to excel. It is the source of an intrinsic motivation and drive to achieve. It relinquishes personal gains in favor of the common good. *Generative conversation* refers to a deep and meaningful dialog to create unity of thought and action. *Systems perspective* is the ability to see things holistically by understanding the interconnectedness of the parts. The foregoing elements give rise to the five core capabilities of learning organizations, namely: personal mastery; shared vision; mental models; team learning and dialog; and systems thinking. These five disciplines are described below. Figure 1 below shows the core capabilities and their relationships.

### Personal Mastery

Senge [30] describes that personal mastery is the cornerstone and ‘spiritual’ foundation of the learning organization. It is born out of a creative orientation and systemic perspective. Personal mastery instils a genuine desire to do well and to serve a noble purpose. People exhibiting high levels of personal mastery focus “on the desired result itself, not the process or the means they assume necessary to achieve that result” [30]. These people can “successfully focus on their ultimate intrinsic desires, not on secondary goals. This is a cornerstone of Personal Mastery”. Personal mastery also requires a commitment to truth, which means to continually challenge “theories of why things are the way they are”. Without committing to the truth, peo-

ple all too quickly revert to old communication routines which can distort reality and prevent them from knowing where they really stand.

### Shared Vision

It is commonly assumed that in contemporary organizations senior management can develop a vision which employees will follow with genuine commitment. This is a fallacy. Simply promoting a ‘vision statement’ could result in a sense of apathy, complacency and resentment. Instead, there needs to be a genuine endeavor to understand what people will commit to. The overriding vision of the group must build on the personal visions of its members. Shared vision should align diverse views and feelings into a unified focus.

This is emphasized by Arie de Geus [9] when he describes what makes a truly extraordinary organization. “The feeling of belonging to an organization and identifying with its achievements is often dismissed as soft. But case histories repeatedly show that a sense of community is essential for long term survival”. For example, when Apple Corporation challenged IBM, it was in its ‘adolescent’ years, characterized by creativity, confidence and even defiance. This is similar to the spirit in Team New Zealand when it competed against the bigger-budget syndicates! Within these organizations there is a real passion for the outcome; a common vision for success [19].

Creating a shared vision is the most fundamental job of a leader [26]. By creating a vision, the leader provides a vehicle for people to develop commitment, a common goal around which people can rally, and a way for people to *feel* successful. The leader must appeal to people's emotions if they are to be energized towards achieving the goal. Emotional acceptance of, and belief in, a vision is far more powerful in energizing team members than is intellectual recognition that the vision is simply a 'good idea'. One of the most powerful ways of communicating a vision is through a leader's personal example and actions, demonstrating behavior that symbolizes and furthers that vision.

### Mental Model and Leadership

Mental models reflect beliefs, assumptions and feelings that shape one's world views and actions. They are formed through family, education, professional and social learning based, on the most part, on cultural and social norms. Mental models, however, can be altered and aligned.

Organizations are often constrained by deep-seated belief systems, resulting in preconceived ideas on how things ought to perform. Goodstein and Burke ([14] p. 10), pioneers in the field of social psychology of organizations, observed that 'the first step in any change process is to unfreeze the present patterns of behavior as a way of managing resistance to change'. The leader has a pivotal role in dismantling negative mental models and shaping new ones.

In order to get people to engage in open discussions of issues that affect the organization, a leader must appeal to their emotions and must get beyond the superficial level of communication. In the 1970s Shell Oil undertook major changes in its leadership approach and communications style. According to a manager at Shell, "When I tried to talk personally about an issue rather than say 'here's the answer', it was powerful. It caused me to engage in dialog with others that resulted in mutual learning on all sides" ([7] p. 71).

The leader is a 'designer', and part of that role is designing the governing ideas of purpose and core values by which people will live [30,31]. In this role, the leader must propose and model the manner in which the group has to operate internally. This provides ample opportunities for leaders to examine their deeply held assumptions about the task, the means to accomplish it, the uniqueness of the people and the kinds of relationship that should be fostered among the people. Only after people have observed and *experienced* the organizational values in practice would these values become the basis for prolonged

group behavior. These values should be manifested first and should be most visible in the leader's own behavior.

Leadership, especially in knowledge-based organizations, must be distributed and shared to a far greater extent than it was in the past. For example, in the Chicago Bulls basketball team, Michael Jordan changed his role: it became not only that of an individually brilliant player but *also* that of a leader whose job it was to raise the level of play of other team members. After this transition, the Bulls began their record run of championship seasons [7].

### Team Learning and Dialog

The word 'dialog' comes from the Greek words *dia* and *logos*. It implies that when people engage in dialog, the meaning *moves through* them – Thus, it enables them to 'see through words' [16]. Dialog is an essential requirement for organizational learning. It results from generative conversation, shared vision, and transparent mental models. Dialog creates a deep sense of listening and suspending one's own views. Feedback is an integral aspect of dialog.

Communication routines in organizations are generally anti-learning and promote mediocrity. They include 'defensive routines' [2] – statement that can stifle dialog and innovative thinking. Exposing and unlearning such routines, and understanding the powerful detrimental impact they have on learning, are serious challenges many organizations face if they are to create effective learning environments.

Many leaders are charismatic and are highly eloquent when it comes to presenting their ideas; that's often why they get to the top of the organization. However, many appear to lack the ability to extract the very best from employees in a non-threatening manner. Without this ability, leaders may miss many good ideas, or might act on many bad ones.

In a group context, encouragement from the leader and mutual encouragement among group members is essential. Furthermore, personal differences must be put aside in order for effective dialog to ensue.

### How Organizations Learn

(This section is edited from Wikipedia: [http://en.wikipedia.org/wiki/Organizational\\_learning](http://en.wikipedia.org/wiki/Organizational_learning))

"Argyris and Schon were the first to propose concepts and models that facilitate organizational learning, the following literatures have followed in the tradition of their work:

- March and Olsen [23] attempt to link up individual and organizational learning. In their model, individual be-

liefs lead to individual action, which in turn may lead to an organizational action and a response from the environment which may induce improved individual beliefs and the cycle then repeats over and over. Learning occurs as better beliefs produce better actions.

- Argyris and Schon [3] distinguish between single-loop and double-loop learning, related to Gregory Bateson's concepts of first and second order learning. In single-loop learning, individuals, groups, or organizations modify their actions according to the difference between expected and obtained outcomes. In double-loop learning, the entities (individuals, groups or organization) question the values, assumptions and policies that led to the actions in the first place; if they are able to view and modify those, then second-order or double-loop learning has taken place. Double loop learning is the learning about single-loop learning.
- Kim [17], as well, in an article titled "The link between individual and organizational learning", integrates Argyris, March and Olsen and another model by Kofman into a single comprehensive model; Further, he analyzes all the possible breakdowns in the information flows in the model, leading to failures in organizational learning; For instance, what happens if an individual action is rejected by the organization for political or other reasons and therefore no organizational action takes place?
- Nonaka and Takeuchi [27] developed a four stage spiral model of organizational learning. They started by differentiating Polanyi's concept of "tacit knowledge" from "explicit knowledge" and describe a process of alternating between the two. Tacit knowledge is personal, context specific, subjective knowledge, whereas explicit knowledge is codified, systematic, formal, and easy to communicate. The tacit knowledge of key personnel within the organization can be made explicit, codified in manuals, and incorporated into new products and processes. This process they called "externalization". The reverse process (from explicit to implicit) they call "internalization" because it involves employees internalizing an organization's formal rules, procedures, and other forms of explicit knowledge. They also use the term "socialization" to denote the sharing of tacit knowledge, and the term "combination" to denote the dissemination of codified knowledge. According to this model, knowledge creation and organizational learning take a path of socialization, externalization, combination, internalization, socialization, externalization, combination... etc. in an infinite spiral.
- Flood [10] discusses the concept of organizational learning from Peter Senge and the origins of the theory from Argyris and Schon. The author aims to "re-think" Senge's *The Fifth Discipline* through systems theory. The author develops the concepts by integrating them with key theorists such as Bertalanffy, Churchman, Beer, Checkland and Ackoff. Conceptualizing organizational learning in terms of structure, process, meaning, ideology and knowledge, the author provides insights into Senge within the context of the philosophy of science and the way in which systems theorists were influenced by twentieth-century advances from the classical assumptions of science.
- Nick Bontis et al. [4] empirically tested a model of organizational learning that encompassed both stocks and flows of knowledge across three levels of analysis: individual, team and organization. Results showed a negative and statistically significant relationship between the misalignment of stocks and flows and organizational performance.
- Imants [15] provides theory development for organizational learning in schools within the context of teachers' professional communities as learning communities, which is compared and contrasted to teaching communities of practice. Detailed with an analysis of the paradoxes for organizational learning in schools, two mechanisms for professional development and organizational learning, (1) steering information about teaching and learning and (2) encouraging interaction among teachers and workers, are defined as critical for effective organizational learning.
- Common [8] discusses the concept of organizational learning in a political environment to improve public policy-making. The author details the initial uncontroversial reception of organizational learning in the public sector and the development of the concept with the learning organization. Definitional problems in applying the concept to public policy are addressed, noting research in UK local government that concludes on the obstacles for organizational learning in the public sector: (1) overemphasis of the individual, (2) resistance to change and politics, (3) social learning is self-limiting, i.e. individualism, and (4) political "blame culture". The concepts of policy learning and policy transfer are then defined with detail on the conditions for realizing organizational learning in the public sector."

### Modeling for Organizational Learning

In general, the *process* of model building can be an effective conduit for collective learning. System Dynamics modeling, in particular, can be used to enhance organizational

learning [35] through rapid feedback and experimentation and its facility to test assumptions and mental models. As we have discussed, dealing effectively with mental models is one of the core competencies for organizational learning

Ackoff [1] likens complex problems to “messes”. “Messy problems are defined as situations in which there are large differences of opinion about the problem or even on the question of whether there is a problem. Messy situations make it difficult for a management team to reach agreement. System Dynamics modeling with groups known as Group Model Building (GMB) is a powerful tool for dealing with these. SD and GMB are especially effective in dealing with semi-structured and ill-structured decision situations.”

GMB offers an opportunity to align and share piecemeal mental models and create the possibility of assimilating and integrating partial mental models into a holistic system description [38,39]. GMB and SD can help uncover ‘illusions’ that may occur due to the fact that the definition of a problem may be a socially constructed phenomenon that has not been put to test ([18] p. 84).

### Learning Laboratory

(This section is adapted from [21], Chapter 6)

Learning laboratory is a setting as well as a process in which a group can learn together. The purpose of the learning lab is to enable managers to test their long held assumptions and to experiment and ‘see’ the consequences of their actions, policies and strategies. This often results in finding inconsistencies and the discovery of *unintended* consequences of actions and decisions, *before* they are implemented. System Dynamics models known as Microworlds or Management Flight Simulators (MFS) are the ‘engine’ behind the learning lab. “Just as an airline uses flight simulators to help pilots learn, system dynamics is, partly, a method for developing management flight simulators, often computer simulation models, to help us learn about dynamic complexity, understand the sources of policy resistance, and design more effective policies.” ([35], p. 4)

A learning lab is distinct from so-called management games. In management games, the players are required to compete – design the ‘best’ strategy and ‘beat’ other players or teams. The competitive nature of management games often encourages aggressive and individualistic behavior with scant regard for group learning and gaining deep insights. The learning lab, in contrast, aims to enhance *learning*: To test individual and group mental models and to provide deeper understanding and insights into why systems behave the way they do. This will help the partic-

ipants to test their theories and discover inconsistencies and ‘blind spots’ in policies and strategies *before* they are implemented.

A significant benefit of the learning lab stems from the process in which participants examine, reveal and test their mental models and those of their organization. The learning lab can also help participants

- To align strategic thinking with operational decisions;
- To connect short-term and long-term measures;
- To facilitate integration within and outside the organization;
- To undertake experimentation and learning;
- To balance competition with collaboration.

### Managerial Practice Field

Team and teamwork are parts of the lexicons of numerous organizations today. Company after company has reorganized work around a variety of team concepts. From factories to hospitals, *titles* like ‘manager’ and ‘supervisor’ have been replaced by *roles* such as ‘facilitator’ and ‘team leader’. Despite this level of attention to team and teamwork the expected benefits have been marginal at best.

But when we examine real teams, such as sporting teams, orchestras or ballet companies more closely, they all share one key characteristic. That is they *practice* a lot more than they ‘perform’. Practice involves allowing time and space to experiment with new ways, try different approaches and most importantly, make mistakes without the fear of failure. In fact, making mistakes is indispensable to learning. One cannot learn from doing things right all the time! Yet a great deal of organizational energy and attention is devoted to the prevention and masking of mistakes.

But, what is the *practice field* for management teams? The fact is that the practice field is, by and large, absent from the managerial world. In other words, there is no time and no space for management to ‘practice’ in the true sense of the word – to experiment, make mistakes and learn together. In this era of restructuring and downsizing, lack of time is the greatest impediment to managerial and organizational learning. As a recent advertisement by IBM reads, “Innovative Thinking! We don’t even have time for bad thinking”. The pace in the modern work environment is so unrelenting that there is virtually no room for managers to slow down, to practice, to reflect and learn. The consequence of this lack of practice and learning space is grave, in that most organizations only achieve a small fraction of their potential – about 5%, according to Jay Forrester, the father of System Dynamics [13].



In order to fill this gap, the concept of learning laboratory has been developed to provide practice fields for managers. The learning lab allows learning to become an integral part of managerial work and helps learning to become institutionalized [17].

### Aligning Mental Models Through the Learning Laboratory

Mental models are formed throughout one's life. Family, school, culture, religion, profession and social norms play important roles in this formation. Therefore, modifying one's mental model is not a small matter. The most effective way to check one's mental models is to *experience* alternative realities at first hand and see their implications with a new 'lens' [5].

There are rarely any opportunities in the course of a manager's daily work for him/her to engage in lengthy, drawn-out experimentation. Learning in a 'laboratory' setting is a viable and powerful alternative. Fortunately, advanced computers and sophisticated system dynamics software have enabled the creation of managerial learning labs where managers can experiment, test their theories and learn rapidly. Thus, learning labs can play a significant role in clarifying and changing mental models. Learning lab deals with mental models at three levels [33], as described below.

- *Mapping* mental models. This step begins at the conceptualization phase. Here, the learning lab participants articulate and clarify their assumptions, views, opinions, and biases regarding the issue at hand.
- *Challenging* mental models. The participants identify and discuss inconsistencies and contradictions in their assumptions. This step will begin at the conceptualizations phase and will continue to the experimentation phase.
- *Improving* mental models. Having conducted experimentation and testing, the participants reflect on the outcomes. This may cause them to alter, adjust, improve and harmonize their mental models.

The laboratory setting provides a neutral and 'safe' space for the participants to create a shared understanding of complex and endemic issues. The following characteristics of the learning lab provide a powerful catalyst for alignment of divergent mental models in the organization.

- The laboratory environment is neutral and non-threatening. The emphasis is on learning and theory building (what we *don't* know), not on winning or display of knowledge.

- Lack of hierarchy. Managers and staff are equal in this environment. The traditional hierarchy is minimized in the laboratory setting.
- The response time is fast. Hence, the feedback cycle is short, which leads to rapid learning.
- There is no cost or 'loss of face' attached to failure. Hence, it is safe to make mistakes. In fact, mistakes provide opportunities for learning.
- People can see the consequences of their actions first hand. No one attempts to convince or teach anyone else or force his or her preconceived views on others. People learn by themselves and through group interactions.

### Implications for Management

The practice field and the learning lab concepts offer fresh and challenging implications for managers and their role. They suggest that a leader/manager should think as a *scientist*, be open to and welcome hard questions, experiment with new ideas, and be prepared to be *wrong*. This requires managers to learn systems thinking skills and use them not just for 'solving' problems but as powerful tools for communication, team building and organizational learning. This means that an effective leader should be the 'designer' of the ship and not its captain [31]. Once they have designed a new structure, strategy, policy or procedure then the managers/leaders should allow (i. e. create a practice field for) the staff to experience the new design, and experiment with it and learn for themselves – the desired outcome is shared understanding leading to alignment of thoughts and actions. This is the essence of organizational learning.

### Future Directions

#### Agent-Based Modeling (ABM)

Agent based modeling (ABM) is an emerging modeling technology which draws its theories and techniques from complexity science [29]. While System Dynamics and Agent-Based Modeling (ABM) use different modeling philosophies and approaches, they can be used complementarily and synergistically.

System Dynamics focuses on modeling structures (i. e. relationships, policies, strategies) that underlie behavior of systems. This may be viewed as a weakness of system dynamics approach in that behavior is assumed to be solely a function of structure (model relationships defined a priori). In contrast, in ABM, organizations are modeled as a system of semi-autonomous decision-making elements – purposeful individuals called *agents*. Each agent individually assesses its situation and makes de-

cisions based upon value hierarchies representing goals, preferences, and standards for behavior. Thus, macro-behavior is not modeled separately but *emerges* from the micro-decisions of individual agents. In other words, in agent based modeling, “emergent” behavior is expected as a result of agents’ interactions. This is a key difference between the two approaches.

While system dynamics acknowledges the critical role of individual and organizational mental models (e. g., motivations, values, norms, biases, etc.) it does not explicitly model them. SD utilizes factual data or “cold knowledge” and does not take into account decision makers ‘mood’. In contrast, ABM attempts to capture “warm knowledge”, representing emotional and human context of decision-making.

Recent advances in video game technology allow the development of multi-agent, artificial ‘society’ simulators with capabilities for modeling physiology, stress and emotion in decision-making [34]. At the simplest level, an agent-based model consists of a system of agents and their relationships. This new approach enables superior understanding of the complexity in organizations and their relevant business environments. This in turn provides an opportunity for new sophistications in game-play that enhances decision-making. Experience with agent-based modeling shows that even a simple agent-based model can exhibit complex behavior patterns and provide valuable information about the dynamics of the real world system that emulates them.

Despite their differences, SD and ABM can be used in a complementary fashion. Both ABM and SD are powerful tools for transforming information into knowledge and understanding leading to individual and group learning. However, the transition from knowledge to understanding may not be immediate or transparent. This requires a deep shift in mental models through experimentation and group learning.

### Systems Thinking and Sustainability

Systems Thinking has a natural affinity with sustainability modeling and management. Sustainability issues are complex; cut across several disciplines; involve multiple stakeholders and require a long term integrated approach. Thus, the systems paradigm and tools have direct and powerful applications in sustainability issues and management.

The applications of system dynamics in sustainability go back to the early 1970s with Jay Forrester’s “World2 and World3 analyzes against 30 years of history”, followed by “World Dynamics” and “Limits to Growth” [24]

and “Beyond the Limits” [25]. “The politics of the environment has also evolved dramatically since 1970. Public awareness of the reality of the environmental challenge has risen; Ministries of Environment have become commonplace” ([28] p. 220). As an example today concern over carbon emissions has already become an international currency. As a result, sustainability has brought a fresh challenge for governments, business and industry, scientists, farmers and all the citizens of the world collectively to find systemic solutions that are mutually and globally agreeable. Systems Thinking and System dynamics can make real and valuable contributions to addressing this challenge.

### Bibliography

#### Primary Literature

1. Ackoff RA (1999) Re-creating the corporation – A design of organizations for the 21st century. Oxford University Press, Oxford
2. Argyris C (1992) The next challenge for TQM: Overcoming organisational defences. *J Qual Particip* 15:26–29
3. Argyris C, Schon D (1978) Organizational learning: A theory of action perspective. Addison-Wesley, Reading
4. Bontis N, Crossan M, Hulland J (2002) Managing an organizational learning system by aligning stocks and flows. *J Manag Stud* 39(4):437–469
5. Brown JS (1991) Research that reinvents the corporation. *Harv Bus Rev* 68:102–111
6. Cavana R, Maani K (2004) A methodological framework for integrating systems thinking and system dynamics. In: *System Dynamics Society Proceedings*. Oxford
7. Cohen E, Tichy N (1997) How leaders develop leaders. Training and Development, May
8. Common R (2004) Organisational learning in a political environment: Improving policy-making in UK government. *Policy Stud* 25(1):35–49
9. De Geus A (1997) The living company. *Harv Bus Rev* 75(2):51–59
10. Flood RL (1999) Rethinking the fifth discipline: Learning within the unknowable. Routledge, London
11. Forrester JW (1961) Industrial dynamics. Productivity Press, Cambridge
12. Forrester JW (1971) World dynamics. Wright-Allen (Subsequently re-published by Productivity Press, and Pegasus Communications)
13. Forrester JW (1994) Building a foundation for tomorrow’s organizations. In: *Systems thinking in action video collection*, vol 1. Pegasus Communications, Cambridge
14. Goodstein L, Burke W (1991) Creating successful organisation change. *Organ Dyn* 19(4):5–17
15. Imants J (2003) Two basic mechanisms for organizational learning in schools. *Europ J Teach Educ* 26(3):293–311
16. Isaacs W (1993) Taking flight: Dialogue, collective thinking and organisational learning. *Organ Dyn* 22(2):24–39
17. Kim DH (1993) The link between individual and organizational learning. *Sloan Manag Rev* 35(1):37–50

18. Maani K (2002) Consensus building through systems thinking – the case of policy and planning in healthcare. *Aust J Inform Syst* 9(2):84–93
  19. Maani K, Benton C (1999) Rapid team learning. Lessons from team New Zealand's America's cup campaign. *Organ Dyn* 27(4)
  20. Maani K, Cavana R (2007) Systems methodology. *Syst Think* 18(8):2–7
  21. Maani K, Cavana R (2007) Systems thinking, system dynamics – Managing change and complexity, 2nd edn. Prentice Hall, Pearson Education, Auckland
  22. Maani K, Pourdehnad J, Sadehi H (2003) Integrating system dynamics and intelligent agent-based modelling – theory and case study. *Euro INFORMS*, Istanbul
  23. March JG, Olsen JP (1975) The uncertainty of the past; Organizational ambiguous learning. *Europ J Polit Res* 3:147–171
  24. Meadows DH, Meadows DL, Randers J, Behren W (1972) The limits to growth. Universe Press, New York
  25. Meadows DH, Meadows DL, Randers J (1992) Beyond the limits. Chelsey Green, Post Mills
  26. Nadler DA, Tushman ML (1990) Beyond the charismatic leader: Leadership and organisational change. *Calif Manag Rev*
  27. Nonaka I, Takeuchi H (1995) The knowledge creating company. Oxford University Press, New York
  28. Randers J (2000) From limits to growth to sustainable development or SD (sustainable development) in a SD (system dynamics) perspective. *Syst Dyn Rev* 16(3):213–224
  29. Rothfeder J (2003) Expert voices: Icosystem's Eric Bonabeau. *CIO Insights*
  30. Senge P (1990) The fifth discipline: The art and practice of the learning organisation. Currency
  31. Senge P (1990) The leader's New Work: Building learning organisation's. *Sloan Manag Rev*:7–23
  32. Senge P (1992) Building learning organisation's. *J Qual Particip*:1–8
  33. Senge P, Sterman JD (1991) Systems thinking and organizational learning: Acting locally and thinking globally in the organization of the future. In: Kochan T, Useem M (eds) *Transforming organizations*. Oxford University Press, Oxford
  34. Silverman BG et al (2002) Using human models to improve the realism of synthetic agents. *Cogn Sci Q* 3
  35. Sterman JD (2000) Business dynamics, systems thinking and modeling for a complex world. McGraw-Hill, Irwin
  36. System Dynamics Society website <http://www.systemdynamics.org/>
  37. US Department of Energy Introduction to system dynamics, A systems approach to understanding complex policy issues, US Department of Energy. <http://www.systemdynamics.org/DL-IntroSysDyn/inside.htm>
  38. Vennix JAM (1995) Building consensus in strategic decision-making: System Dynamics As A Support System. *Group Decis Negot* 4(4):335–355
  39. Vennix JAM (1996) Group model-building: Facilitating team learning using system dynamics. Wiley, Chichester, chapt 5
  40. Wolstenholme E (1997) System dynamics in the elevator (SD1163), e-mail communication, 24 Oct 1997 [systemdynamics@world.std.com](mailto:systemdynamics@world.std.com)
- Dynamics Group", Sloan School of Management, MIT)
- Industrial and Economic Dynamics: The Foundations**
- Forrester JW (1961) *Industrial dynamics*. Productivity Press, Cambridge (Presents dynamic analysis of a business problem through a model of a production-distribution system that shows oscillatory behavior. Policies to improve system performance are discussed, and numerous policy experiments are demonstrated. Includes full equation listing.)
- Forrester JW (1968) *Principles of systems*. Productivity Press, Cambridge (System structure and behavior are differentiated, with examples showing how structure determines behavior. Rates and levels are described. Inventory model shows effects of delivery delay and resulting production cycles.)
- Forrester JW (1975) *Collected papers of Jay W. Forrester*. Productivity Press, Cambridge (Includes many seminal papers, such as *Industrial Dynamics: A Major Breakthrough for Decision Makers*; *Common Foundations Underlying Engineering and Management*; *A New Corporate Design*; *Market Growth as Influenced by Capital Investment*; and *Counterintuitive Behavior of Social Systems*.)
- Forrester JW (1989) The beginnings of system dynamics (Working Paper No. D-4165). System Dynamics Group, Sloan School of Management, MIT, Cambridge (A personal history beginning on the high plains of western Nebraska. Describes the early projects that shaped the field.)
- Mass NJ (1975) *Economic cycles: An analysis of underlying causes*. Productivity Press, Cambridge (Shows how production scheduling and work force management policies generate the 3–5 year business cycle. Economic cycles, in turn, are caused by capital investment policies that fail to account for delays in acquiring long-lead time plant and equipment.)
- Meadows DL (1970) *Dynamics of commodity production cycles*. Productivity Press, Cambridge (Develops a simple generic model of commodity supply and demand with explicit production capacity and delays, prices and markets. Applies the model to hogs, chicken and cattle.)
- Urban and Public Policy Dynamics**
- Alfeld LE, Graham AK (1976) *Introduction to urban dynamics*. Productivity Press, Cambridge (A very readable introductory text. Uses the urban system as an example to teach general points about modeling methods, formulation and analysis.)
- Forrester JW (1969) *Urban dynamics*. Productivity Press, Cambridge (Seminal model of urban growth and decay, controversial then and vindicated now. Chapter 6 describes general characteristics of complex systems such as compensating feedback and shifting the burden to the intervener.)
- Mass NJ (ed) (1974) *Readings in urban dynamics*, vol I. Productivity Press, Cambridge (Extensions, modification, and responses to criticisms of the Urban Dynamics model.)
- Schroeder WW, III, Sweeney RE, Alfeld LE (eds) (1975) *Readings in urban dynamics*, vol II. Productivity Press, Cambridge (Further extends and explores the Urban Dynamics model.)
- Limits to Growth and Other Global Models**
- Forrester JW (1973) *World Dynamics*, 2nd edn. Productivity Press, Cambridge (The first global model, on which Limits to Growth was based. The extreme simplicity of the model allowed it to be presented to a wide audience.)
- Meadows DL, Meadows DH (eds) (1974) *Toward global equilibrium: Collected papers*. Productivity Press, Cambridge (Describes and explores, through system dynamics models, poli-

## Books and Reviews

(This section is due to M. Anjali Sastry and John D. Sterman, "An Annotated Survey of the Essential System Dynamics Literature System

cies for sustainability designed to avoid the collapse shown in the 'business as usual' WORLD3 scenarios.)

Meadows DH, Meadows DL, Randers J, Behrens WW III (1972) *The limits to growth: A report for the club of Rome's project on the predicament of mankind*. Universe Books, New York (Classic controversial study of the human future. Nontechnical presentation of structure, assumptions, and results of the WORLD3 model. Concluded that present policies were unsustainable; shows how alternate policies could stabilize population at a high standard of living.)

Meadows DL, Behrens WW III, Meadows DH, Naill RF, Randers J, Zahn EKO (1974) *Dynamics of growth in a finite world*. Productivity Press, Cambridge (Full documentation and data for the WORLD3 model used in the *Limits to Growth*. Describes the structure and assumptions; includes all data needed for complete replication of all runs in the popular book. Formulations described here may be useful to all system dynamics modelers.)

Meadows D, Richardson J, Bruckmann G (1982) *Groping in the dark*. Wiley, New York (Describes a range of global models built under different approaches and discusses the strengths, weaknesses, and implications of each. Presented in an engaging, personal style.)

Meadows DH, Meadows DL, Randers J (1992) *Beyond the limits: Confronting global collapse, envisioning a sustainable future*. Chelsea Green, Post Mills (Follows up on *Limits to Growth*. Shows that many problems described in 1972 have worsened, as predicted by the model. Argues for a shift in values necessary to create a sustainable and equitable future.)

#### **SD for Management: Firm and Market Models**

Coyle RG (1977) *Management System Dynamics*. Wiley, New York (Text emphasizing managerial modeling, with a focus on operations and examples including discrete elements.)

Hall RI (1976) A system pathology of an organization: The rise and fall of the Old Saturday Evening Post. *Adm Sci Q* 21(2):185–211 (A case-study using a system dynamics model to explain how failure to understand the feedbacks among policies governing ad rates, ad and editorial pages, marketing, and pricing lead to the failure of the Post just as circulation reached an all-time high.)

Lyneis JM (1980) *Corporate planning and policy design*. Productivity Press, Cambridge (Begins with a simple model of inventory management in a manufacturing firm and gradually extends the model to one of the entire firm.)

Merten PP (1991) Loop-based strategic decision support systems. *Strat Manag J* 12:371–382 (Describes a model of a multinational firm establishing new markets in less-developed countries. Captures qualitative shifts in firm structure and organization endogenously as the firm evolves.)

Morecroft JDW (1984) Strategy support models. *Strat Manag J* 5(3):215–229 (Describes the use of models as participants in the ongoing dialogue among managers regarding strategy formation and evaluation. Emphasizes the processes for model development and use that enhance the utility of modeling in design of high-level corporate strategy.)

Morecroft JDW, Lane DC, Viita PS (1991) Modelling growth strategy in a biotechnology startup firm. *Syst Dyn Rev* 7(2):93–116 (Describes a case-study of a start-up in which system dynamics modeling helps to define a desirable growth strategy for the firm. The integrated model generated strategies

that allowed different parts of the firm to choose consistent approaches.)

Roberts EB (ed) (1978) *Managerial applications of system dynamics*. Productivity Press, Cambridge (Extensive collection of early corporate models, including history and commentary by practitioners. Covers R&D management, production and operations, human resources, and other applications areas.)

#### **Economic Models**

Forrester JW (1989) The system dynamics national model: Macrobbehavior from microstructure. In: Milling PM, Zahn EOK (eds) *Computer-based management of complex systems: International System Dynamics Conference*. Springer, Berlin (Provides an overview of the national modeling project in which both micro- and macro-economic factors are included. Model generates business cycles, inflation, stagflation, the economic long wave, and growth.)

Saeed K (1986) The dynamics of economic growth and political instability in the developing countries. *Syst Dyn Rev* 2(1):20–35 (Shows how rapid economic development can generate social and political instability through a model that links socio-political factors to economic development.)

Sterman JD (1985) A behavioral model of the economic long wave. *J Econ Behav Organ* 6(1):17–53 (Proposes and tests a simple model of the long wave. The intended rationality of each decision rule is tested and the long wave is explained as the unintended result of the interaction of locally rational decision processes. The model is the basis for the STRATAGEM-2 game, and can exhibit chaos.)

Sterman JD (1989) Deterministic chaos in an experimental economic system. *J Econ Behav Organ* 12:1–28 (Sterman's 1985 model of the long wave is converted into a management flight simulator and used as an experiment in which subjects make the capital investment decision. Simple decision rules capturing subject's policies are estimated and explain their behavior well. Simulation of these rules yields deterministic chaos for about 25% of the subjects.)

Sterman JD (1986) The economic long wave: Theory and evidence. *Syst Dyn Rev* 2(2):87–125 (Comprehensive overview of the theory of long waves arising from the System Dynamics National Model. Reviews the feedback structures responsible for the long wave and empirical evidence supporting the dynamic hypotheses. Discusses the role of innovation and political value change.)

#### **Conceptualizing, Formulating and Validating Models**

Barlas Y (1989) Multiple tests for validation of system dynamics type of simulation models. *Europ J Operat Res* 42(1):59–87 (Discusses a variety of tests to validate SD models, including structural and statistical tests.)

Barlas Y, Carpenter S (1990) Philosophical roots of model validation: Two paradigms. *Syst Dyn Rev* 6(2):148–166 (Contrasts the system dynamics approach to validity with the traditional, logical empiricist view of science. Finds that the relativist philosophy is consistent with SD and discusses the practical implications for modelers and their critics.)

Forrester JW (1980) Information sources for modeling the national economy. *J Am Stat Assoc* 75(371):555–574 (Argues that modeling the dynamics of firms, industries, or the economy requires use of multiple data sources, not just numerical data and statistical techniques. Stresses the role of the mental and descriptive data base; emphasizes the need for first-hand field study of decision making.)



- Forrester JW (1985) The model versus a modeling process. *Syst Dyn Rev* 1(1):133–134 (The value of a model lies not in its predictive ability alone but primarily in the learning generated during the modeling process.)
- Forrester JW (1987) Fourteen 'Obvious Truths'. *Syst Dyn Rev* 3(2):156–159 (The core of the system dynamics paradigm, as seen by the founder of the field.)
- Forrester JW (1987) Nonlinearity in high-order models of social systems. *Europ J Operat Res* 30(2):104–109 (Nonlinearity is pervasive, unavoidable, and essential to the functioning of natural and human systems. Modeling methods must embrace nonlinearity to yield realistic and useful models. Linear and nearly-linear methods are likely to obscure understanding or lead to erroneous conclusions.)
- Homer JB (1983) Partial-model testing as a validation tool for system dynamics. In: *International System Dynamics Conference*, pp 920–932 (How model validity can be improved through partial model testing when data for the full model are lacking.)
- Legasto AA Jr, Forrester JW, Lyneis JM (eds) (1980) *System dynamics*. In: *TIMS studies in the management sciences*, vol 14. North-Holland, Amsterdam (Collection of papers focused on methodology. Includes Forrester and Senge on Tests for Building Confidence in System Dynamics Models and Gardiner & Ford's discussion on Which Policy Run is Best, and Who Says So?)
- Mass N (1991) Diagnosing surprise model behavior: A tool for evolving behavioral and policy insights. *Syst Dyn Rev* 7(1):68–86 (Guidelines for learning from surprise model behavior with tests to resolve anomalous behavior.)
- Morecroft JDW (1982) A critical review of diagramming tools for conceptualizing feedback system models. *Dynamica* 8(1):20–29 (Critiques causal-loop diagrams and proposes subsystem and policy structure diagrams as superior tools for representing the structure of decisions in feedback models.)
- Randers J (ed) (1980) *Elements of the system dynamics method*. Productivity Press, Cambridge (Includes Mass on Stock and Flow Variables and the Dynamics of Supply and Demand; Mass & Senge on Alternative Tests for Selecting Model Variables; and Randers' very useful Guidelines for Model Conceptualization.)
- Richardson GP (1986) Problems with causal-loop diagrams. *Syst Dyn Rev* 2(2):158–170 (Causal-loop diagrams cannot show stock-and-flow structure explicitly and can obscure important dynamics. Offers guidelines for proper use and interpretation of CLDs.)
- Richardson GP, Pugh AL III (1981) *Introduction to system dynamics modeling with DYNAMO*. Productivity Press, Cambridge (Introductory text with excellent treatment of conceptualization, stocks and flows, formulation, and analysis. A good way to learn the DYNAMO simulation language as well.)
- Roberts N, Andersen DF, Deal RM, Grant MS, Shaffer WA (1983) *Introduction to computer simulation: A system dynamics modeling approach*. Addison-Wesley, Reading (Easy-to-understand introductory text, complete with exercises.)
- Sterman JD (1984) Appropriate Summary Statistics for Evaluating the Historical Fit of System Dynamics Models. *Dynamica* 10(2):51–66 (Describes the use of rigorous statistical tools for establishing model validity. Shows how Theil statistics can be used to assess goodness-of-fit in dynamic models.)
- Wolstenholme EF (1990) *System enquiry – A system dynamics approach*. Wiley, Chichester (Describes a research methodology for building a system dynamics analysis. Emphasizes causal-loop diagramming, mapping of mental models, and other tools for qualitative system dynamics.)
- Modeling for Learning: Systems Thinking and Organizational Learning**
- Kim D (1989) Learning laboratories: Designing a reflective learning environment. In: Milling PM, Zahn EOK (eds) *Computer-based management of complex systems: International system dynamics conference*. Springer, Berlin (A case-study of a process designed to convey dynamic insights to participants in a workshop setting designed around a management flight simulator game.)
- Morecroft JDW (1988) System dynamics and microworlds for policymakers. *Europ J Operat Res* 35(3):301–320 (Describes the model-building tools available to managers and policymakers.)
- Morecroft JDW, Sterman JD (eds) (1992) *Modelling for Learning*. *Eur J Operat Res Special Issue* 59(1) (17 papers describing models and methods to enhance learning, both for individuals and organizations. Covers elicitation and group process techniques, management flight simulators, and tools for capturing, representing, and simulating mental and formal models.)
- Richmond B (1990) Systems thinking: A critical set of critical thinking skills for the 90's and beyond. In: Andersen DF, Richardson GP, Sterman JD (eds) *International System Dynamics Conference, 1990* (Proposes a process and skill set to teach systems thinking. The process relies on learner-directed learning. The skill set includes general scientific reasoning and SD, supported by simulation.)
- Senge PM (1990) Catalyzing systems thinking within organizations. In: Masarik F (ed) *Advances in organization development*. Ablex, Norwood (Presents a case study in which the use of system dynamics generated insights into a chronic business problem. Steps in generating, testing and disseminating a system dynamics model are described.)
- Senge PM (1990) *The fifth discipline: The art and practice of the learning organization*. Doubleday Currency, New York (Introduces systems thinking as part of a wider approach to organizational learning. Conveys basic system structures to a non-technical business audience by means of anecdotes and archetypes.)
- Decision Making**
- Morecroft JDW (1983) System dynamics: Portraying bounded rationality. *Omega* 11(2):131–142 (SD models represent decision making as boundedly rational. Reviews and contrasts the concept of bounded rationality as developed by Herbert Simon. Uses Forrester's Market Growth model to show how locally rational decision rules can interact to yield globally dysfunctional outcomes.)
- Morecroft JDW (1985) Rationality in the Analysis of Behavioral Simulation Models. *Manag Sci* 31(7):900–916 (Shows how the intended rationality of decision rules in SD models can be assessed, and how one analyzes a simulation model and output to understand the assumed bounds on rationality in dynamic models. A model of salesforce effort allocation is used to illustrate.)
- Sterman JD (1987) Expectation formation in behavioral simulation models. *Behav Sci* 32:190–211 (Proposes and tests a simple dynamic model of expectation formation in dynamic models (the TREND function). Shows how the TREND function explains a forty year history of inflation forecasts and several different types of long-term energy demand forecasts.)



- Sterman JD (1989) Misperceptions of feedback in dynamic decision making. *Organ Behav Hum Decis Process* 43(3):301–335 (Describes an experiment with a simple economic system in which subjects systematically generate costly oscillations. Estimates decision rules to characterize subject behavior. Finds that people systematically ignore feedbacks, time delays, accumulations, and nonlinearities. These misperceptions of feedback lead to poor quality decisions when dynamic complexity is high.)
- Sterman JD (1989) Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Manag Sci* 35(3):321–339 (Analyzes the results of the Beer Distribution Game. Misperceptions of feedback are found to cause poor performance in the beer game, as in other experiments. Estimates of the subjects' decision rules show they ignore time delays, accumulations, feedbacks, and nonlinearities.)
- Selected Applications of SD**
- Abdel-Hamid TK, Madnick SE (1991) Software project dynamics: An integrated approach. Prentice Hall, Englewood Cliffs (Integrated SD model of the software development process. The model covers design, coding, reviewing, and quality assurance; these are integrated with resource planning, scheduling, and management of software projects. Includes full documentation, validation, and policy tests.)
- Cooper KG (1980) Naval ship production: A claim settled and a framework built. *Interfaces* 10(6) (An SD model was used to quantify the causes of cost overruns in a large military shipbuilding project. One of the first and most successful applications of system dynamics to large-scale project management; initiated a long line of related project modeling work.)
- Ford A, Bull M (1989) Using system dynamics for conservation policy analysis in the Pacific Northwest. *Syst Dyn Rev* 5(1):1–15 (Describes the use of an extensive SD model of electric power generation with endogenous demand. The model is used to evaluate strategies for conservation and new generation capacity. Includes discussion of implementation and integration of the SD model with other existing planning tools.)
- Gardiner LK, Shreckengost RC (1987) A system dynamics model for estimating heroin imports into the United States. *Syst Dyn Rev* 3(1):8–27 (Describes how the CIA used SD to estimate the illegal importation of drugs to the US.)
- Homer JB (1985) Worker burnout: A dynamic model with implications for prevention and control. *Syst Dyn Rev* 1(1):42–62 (Explains how knowledge workers can experience cycles of burnout through a simple system dynamics model. Avoiding burnout requires that one work at less than maximum capacity.)
- Homer JB (1987) A diffusion model with application to evolving medical technologies. *Technol Forecast Soc Chang* 31(3):197–218 (Presents a generic model of the diffusion of new medical technologies. Case studies of the cardiac pacemaker and an antibiotic illustrate how the same model can explain the different diffusion dynamics of successful and unsuccessful technologies.)
- Homer JB (1993) A system dynamics model of national cocaine prevalence. *Syst Dyn Rev* 9(1):49–78 (An excellent model of the interacting dynamics of addiction, policy-setting, and enforcement.)
- Jensen KS, Mosekilde E, Holstein-Rathlou N (1985) Self-sustained oscillations and chaotic behaviour in kidney pressure regulation. In: Prigogine I, Sengler M (eds) *Laws of nature and human conduct*. Taskforce of Research Information and Study on Science, Brussels (Presents a system dynamics model of the dynamics of rat kidneys. Experimental data show previously unexplained oscillations, sometimes chaotic. The model explains how these fluctuations arise. Excellent example of SD applied to physiology.)
- Levin G, Hirsch GB, Roberts EB (1975) The persistent poppy: A computer-aided search for heroin policy. Ballinger, Cambridge (Examines the interactions within a community among drug users, the police and justice system, treatment agencies, and the citizens. Analyzes policies designed to restore the community's health.)
- Levin G, Roberts EB, Hirsch GB, Kligler DS, Roberts N, Wilder JF (1976) *The Dynamics of Human Service Delivery*. Ballinger, Cambridge (Presents a generic theory of human service delivery, with case studies and examples drawn from mental health care, dental planning, elementary education, and outpatient care.)
- Naill RF (1992) A system dynamics model for national energy policy planning. *Syst Dyn Rev* 8(1):1–19
- Naill RF, Belanger S, Klinger A, Peterson E (1992) An analysis of the cost effectiveness of US energy policies to mitigate global warming. *Syst Dyn Rev* 8(2):111–128 (Reviews the 20 year history of the SD energy models used by the US Dept. of Energy to forecast and analyze policy options for national energy security, including the impact of US policies on global climate change.)
- Sklar Reichelt K (1990) Halter marine: A case study of the dangers of litigation. (Working Paper No. D-4179). System Dynamics Group, Sloan School of Management, MIT, Cambridge (A case-study illustrating the use of system dynamics in litigation. Suitable for classroom teaching.)
- Sturis J, Polonsky KS, Mosekilde E, Van Cauter E (1991) Computer model for mechanisms underlying ultradian oscillations of insulin and glucose. *Am J Physiol* 260(Endocrinol. Metab. 23):E801–E809 (New experimental data show that the human glucose/insulin system is inherently oscillatory. An SD model explains these dynamics. The model is validated against detailed physiological data.)
- Cross-Fertilization and Comparative Methodology**
- Allen PM (1988) Dynamic models of evolving systems. *Syst Dyn Rev* 4(1–2):109–130 (Reviews approaches to nonlinear dynamics, self-organization, and evolution developed in the Brussels school by Prigogine, Allen, and others. Provides illustrations and examples.)
- Kim DH (1990) Toward learning organizations: Integrating total quality control and systems thinking. (Working Paper No. D-4036). System Dynamics Group, Sloan School of Management, MIT, Cambridge (Argues that SD and Total Quality Management are complementary approaches to improvement and organizational learning. Systems thinking and modeling are needed to speed the improvement cycle for processes with long time delays.)
- Meadows DH, Robinson JM (1985) *The electronic oracle: Computer models and social decisions*. Wiley (Comparative assessment of the underlying assumptions, boundary, limitations, and uses of different models, including optimization, simulation, and econometrics. Offers guidelines for assessing model assumptions, including ways to recognize the implicit biases of each modeling paradigm.)
- Powers WT (1990) Control theory: A model of organisms. *Syst Dyn*

Rev 6(1):1–20 (An explicit feedback control perspective on perception and decision making in living organisms. Argues the behaviorist and cognitive paradigms have fundamentally misunderstood the concept of feedback. For Powers, feedback allows organisms to control perceptions by altering behavior.)

Radzicki MJ (1990) *Methodologia oeconomiae et systematis dynamis*. *Syst Dyn Rev* 6(2):123–147 (Surveys the institutionalist paradigm in economics and argues that system dynamics is compatible with the institutionalist perspective. The SD approach offers a means by which institutional theories can be formalized and tested.)

Sterman JD (1985) The growth of knowledge: Testing a theory of scientific revolutions with a formal model. *Technol Forecast Soc Chang* 28(2):93–122 (Presents a formal dynamic model of TS Kuhn's theory of scientific revolutions.)

Sterman JD (1988) A skeptic's guide to computer models. In: Grant L, Lanham MD (eds) *Foresight and national decisions*. University Press of America (Reviews different modeling methods and their underlying assumptions in nontechnical language. Provides a list of questions model users should ask to assess whether a model or method are appropriate to the problem.)

#### Other Themes: Pulling the Threads Together

Cooper K, Steinhurst W (eds) (1992) *The system dynamics society bibliography*. System Dynamics Society. Available from Julie Pugh, 49 Bedford Rd., Lincoln MA, USA 01773. (Lists over 3,000 system dynamics journal articles, books, conference proceedings and working papers. Available in computer-readable format and compatible with bibliographic software)

Meadows DH (1989) System dynamics meets the press. *Syst Dyn Rev* 5(1):68–80 (Reviews the history of encounters between SD and the media. Offers guidelines for effective communication to the public at large. Stresses the importance of communicating even the simplest system concepts.)

Meadows DH (1991) *The global citizen*. Island Press, Washington (A collection of Dana's syndicated newspaper columns applying system dynamics principles to problems of everyday life, from organic farming to the fall of the Soviet Union. Emphasizes environmental issues.)

Richardson GP (1991) *Feedback thought in social science*. University of Pennsylvania Press (Traces the history of the concept of feedback in the social sciences through two threads of thought – the cybernetic and feedback threads. System dynamics is placed in context in a readable and scholarly manner.)

#### Software

DYNAMO. Pugh-Roberts Associates, Cambridge MA. (The first widely-used computer language developed to simulate system dynamics models, DYNAMO is still in use, available for mainframes and PCs. Many of the models in the system dynamics literature were simulated in DYNAMO)

DYSMAP. University of Salford, UK (PC-based simulation language with syntax similar to DYNAMO. Includes optimization capability based on hill-climbing.)

Microworld Creator and S<sup>4</sup>. Microworlds Inc., Cambridge MA (Easy to use environment for simulation and gaming. S<sup>4</sup>, the 'industrial strength' version, supports arrays and includes diagnostics for analyzing behavior. Both Creator and S<sup>4</sup> support user-defined information displays and facilitate rapid development of management flight simulators.)

STELLA and ithink. High Performance Systems, Hanover NH. (User-friendly modeling software with full graphical interface. Mod-

els are entered graphically, at the level of the stock and flow diagram. Widely used in education from elementary school up; also used in research and practice.)

Vensim. Ventana Systems, Harvard MA. (Powerful simulation environment for SD models. Runs on workstations and PCs. Includes array capability and a wide range of features for analyzing model behavior.)

## System Dynamics Philosophical Background and Underpinnings

CAMILO OLAYA

Universidad de Los Andes, Bogotá, Colombia

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[System Dynamics](#)

[“Real” World and Presentationalism](#)

[The Discussions on Positivism:](#)

[Presentationalism and Knowledge](#)

[A Brief Note Regarding Social Theory](#)

[Simulation and Method](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Philosophy** The reflection and study of our most basic assumptions – or the assumptions themselves.

**Mental model** A mental image of selected concepts and relationships of the world around us which we consider relevant for explaining the behavior of a particular system.

**Presentationalism** Synonymous of idealism. The view that material objects or external realities do not exist apart from our knowledge or consciousness of them.

### Definition of the Subject

We all tend to take things for granted. Indeed it is a common place to judge formal models exclusively based on the technical grounds and on the logic with which those models were built without a proper reflection on the assumptions underlying those models. This omission is even more pressing in complexity and system science, since these areas represent a novel challenge for philosophers of science – e. g. see an overview in [34].

What is the idea of reality with which we work? What do we assume about human nature? What kind of knowledge do we pursue? What kind of knowledge do we obtain?

What is the scope of rational inquiry? What are the basis and the implications of our own reasoning methods? The identification of how philosophy has shaped the work of scientists – on a conscious or unconscious level – is essential for comprehending the implications, the limitations, and the scope of our very scientific practice. The lack of concern by scientists for these issues may explain many of their failures which has produced just a sort of inertial blindness that is easy to recognize in current scientific debates.

One of the strengths of system dynamics (abbreviated, SD) is that it leads us to make explicit our assumptions about the systems we deal with. This attitude, i. e. the importance of reflection upon our own assumptions, is also fundamental for the very development and practice of system dynamics. Many of the debates on different issues of every day scientific practice such as model conceptualization, formal model building, validation, policy design, etc. are informed and can be enlightened by the reflection on the philosophical background behind those processes. There are also various fundamental aspects of SD that are yet to be demarcated, e. g. the characterization of SD explanations. Furthermore, the ambiguity of the discussions found in large part of related literature, characterized by superficiality, confusion of terms, misdirected arguments, etc. only adds noise and it complicates the advance of a discipline. This article sketches and overview on some basic assumptions regarding the development and the practice of system dynamics. Various suggestions that help to integrate various debates are introduced and important clarifications are also indicated.

## Introduction

The philosophical background and underpinnings of a discipline should have to do with its most basic universal assumptions. Such a discussion becomes difficult if we bear in mind that those assumptions are not necessarily shared by practitioners and researchers. Nevertheless, central premises can be identified which in turn can be related to important questions regarding philosophical concerns such as reality and knowledge.

The article is organized as follows. After this short introduction, the second sections develops an overview of the origins of system dynamics underlining fundamental aspects that formed what can be called the core of the discipline. This historical review highlights the initial interest, purposes and initial assumptions around the foundation of SD. With these elements the following sections introduce various philosophical issues that can be identified underlying system dynamics. Perhaps the central aspect is

presentationalism, a stance associated with the notion of “mental models” which is central in SD; this is the topic of the fourth section. The following section makes a clarification on the controversial issue of positivism and relates presentationalism with knowledge. The sixth section summarizes the position of system dynamics regarding social theory. The seventh section presents the inquiry of explanation clarifying that in spite that SD involves causal models the nature of its kind of explanation can be found in the notion of mechanism. The eighth section introduces the implication of the use of computer simulation as a distinctive epistemology which is different from the traditional discourse in philosophy of science. The ninth section outlines future directions.

Before starting, a brief warning should be made: given the scope of this review and the limited space for covering very wide subjects then this article should be viewed as a broad introductory overview of the different topics. The cited literature has been selected as possible starting points for further inquiry.

## System Dynamics

In order to address a “philosophical background” the first question naturally would be: What is system dynamics? Already this inquiry can be a matter of debate, e. g. [102]. Indeed SD has been labeled as a theory [23,49], a method [18,56,63,98,108], a methodology [81], a field of study [17,78], a tool [61], a paradigm, among other nouns. A natural starting point is the work of Jay Forrester, the founder of system dynamics. A brief historical review should help to grasp the very core we are looking for since it can show the initial motivations, assumptions, and purposes behind the development of SD.

## Genesis of System Dynamics

Jay W. Forrester, member of the Sloan School of Management at MIT, was looking for linkages between engineering and management education given his background in feedback control systems and computers [28]. In 1956 he wrote a “note” to the Faculty Research Seminar, the first ever MIT “D-memo”; in this communication he sketched the worldview of what would be known as “system dynamics” [32].

He started with a strong criticism of economic models. The following were the central aspects: (i) their failure to reflect adequately the loop structures that make up economic systems; in particular this neglect leads to exclude inherent properties of closed loops such as resistance to change, accumulations and delays; (ii) The incapacity for including flows of goods, money, information, and la-

bor, in one single interrelated model; (iii) The exclusion of changing mental attitudes that affect and explain economic processes; (iv) The use of linear equations for describing systems; (v) The restriction of building models constrained by the capacity for manipulating numerical data and solving the equations; (vi) The overconfidence in multiple regression analysis for obtaining coefficients for equations that define economic behavior; (vii) The lack of reflection and discussion on the very assumptions underlying every model preferring an emphasis on the logic with which the model is developed.

After delineating these points, Forrester then proceeded in the same note to highlight techniques that were largely underused at that time: servomechanisms, differential equations, and what he called “the art of simulation”. Anchored on the mentioned assessment and on these developments Forrester conceived “a new avenue of attack for understanding the firm and the economy” ([28] p. 336) envisaging a *new kind of models* that would include aspects such as:

*Dynamic structure:* Detailed attention to the *sequences of actions* which occur in the system being studied and to the *forces* which trigger or temper such actions, with a particular concern on the controlling influences of lags and delays.

*Information flows:* Explicit recognition of information flow channels and information transformation with time and transmission.

*Decision criteria:* Re-examination of the proper decision criteria which must not be defined as depending only on current values of gross economic variables; instead, such criteria must be traced to the motivations, hopes, objectives and optimism of the people involved, including as well what he calls business man’s intuition which “represents a disordered accumulation of basic insights into how people and social systems react. The hope for the future lies in generating an orderly arrangement of basic insights” ([28] p. 342).

*Non-linear systems:* Economic systems present most – if not all – of the time highly non-linear characteristics.

*Differential equations:* The behavior of economic systems should be better described by non-linear differential equations since they have been developed to describe delays, momentum, elasticity, reservoirs, and accelerations, which are better suited quantities for describing the economic world. In practice these equations would be handled as incremental difference equations in order to obtain numerical solutions.

*Incremental changes in variables:* To prefer the formulation of a model in terms of the motivations that cause incremental changes in a variable since the new value of

a variable “can be found by solving the equations for its incremental change and then adding the change to the preceding full value of the variable” ([28] p. 344).

*Model complexity:* Much complex and complete models can be developed with these techniques.

*Empirical solutions:* It is useless to look for explicit unique or “correct” solutions; instead, these models provide diverse solutions according to the different assumptions about the model structure and the initial values of the variables.

*Symbolism and correspondence with real counterparts:* The possibility of having a pictorial representation – a flow diagram – whose processes of information, money, goods, and people, are moved, i. e. simulated, time-step-by-time-step from place to place.

*Structure over coefficient accuracy:* To prefer a structure in which we have confidence using intuitively estimated coefficients instead of using unlikely structures with accurately derived coefficients from statistical data.

A subsequent advance came in 1958 with an article entitled “Industrial Dynamics: A Major Breakthrough for Decision Makers” published in the *Harvard Business Review* [27]. In this article Forrester shaped the previous ideas with the concern that management should evolve from a highly fragmentized art to a profession capable of recognizing unified systems given that the task of management is to interrelate the flows of information, materials, labor, money, and capital equipment. He again emphasized features such as electronic data processing, decision making, simulation, feedback control, and information flows. These elements were presented as the cornerstones of the innovative industrial dynamics program at MIT.

## Industrial Dynamics

The definitive breakthrough came in 1961 with his *magnus opus* “Industrial Dynamics” [24]. The main motivation behind was the development of a *science* for designing and controlling industrial systems, the quest for a management science. In particular he conceived it as “a method of system analysis for management.” (p. 9). He stated:

Industrial dynamics is the study of the information-feedback characteristics of industrial activity to show how organizational structure, amplification (in policies), and time delays (in decisions and actions) interact to influence the success of the enterprise (p. 13).

Forrester underlined four pillars for this new science: information-feedback control theory anchored on the con-



cept of servomechanisms, the study of decision-making processes, an experimental approach to system analysis based on simulation, and the use of computers.

Coming from engineering, he had in mind models that deal with nonlinear dynamic systems whose purpose is to *design* new systems – as opposed to just *explain* systems; the models should show how changes in policies or structure will produce better or worse behavior. In order to accomplish this aim Forrester indicated that we should focus on understanding the characteristics of the system in hands (instead of looking for specific predictions) and on our assumptions about them. “We then have a means for tracing the implications of our assumptions” ([24] p. 55).

What should be included in a model? Forrester underlined that “there will be no such thing as *the* model of a social system, any more than there is *the* model of an aircraft ... the factors that must be included arise directly from the questions that are to be answered” ([24] p. 60). It is expected that these factors will include closed-loop information-feedback structures that give rise to so much of the interesting behavior. An important aspect of this new kind of models is symbolism and pictorial representations by means of flow diagrams with a special emphasis on correspondence: the model variables should correspond to those in the system being represented. In this book Forrester also demarcated the network structure of this new kind of models as made of four basic components: accumulations (levels), flows (rates), decision functions, and information channels; these networks trace cause-effect relationships which are described via mathematical formalization. He also discussed in detail how to represent delays and how to model decision processes which are particularly defined by general policies, i. e. rules that state how operation decisions are made converting information into action; this study of general policies explains the importance for SD of the examination of human decision-making processes. Another fundamental characteristics are continuous flows and aggregation: “grouping of individual events into classes ... Our interest in the model ... is from the viewpoint above the separate individual transactions” ([24] p. 65); it is assumed that different individual items are controlled by the same identical decision-function; this notion leads to aggregation which is a distinctive aspect of SD: “items controlled by sufficiently similar policies that depend on sufficiently equivalent information sources may be combined into a single channel” ([24] p. 109); the central criterion for such aggregation is the purpose of the model. Finally, model significance (or validity) rests on its suitability for a particular purpose which is motivated by the design of improved industrial and economic systems.

## Principles of System Dynamics

The main concern in these initial steps were business operations. However, Forrester sketched a glance to a broader view in the final part of his book; there, he speaks of “system” dynamics since “the study of systems can provide a framework to unite subjects ... The dynamic model represents a system as broad as one chooses to describe” ([24] pp. 344, 346). Indeed, looking for a more general view he presents a section of “principles of systems structure”:

The principles to be discussed here all arise in the context of information-feedback systems. They are systems principles. They are not the principles of the management art such as have been taught in organization, production, and human relations courses. Because the principles apply to systems behavior, they do not fall into neat separate packages ... The concept of a system implies interaction and interdependence. In attempting to identify factors that are common to all systems, we must keep the essential indivisibility in mind ([24] pp. 347, 348).

Indeed, Forrester reaffirmed this general *systems* view in a follow-up book entitled *Principles of Systems* [25] published in 1968. He gives a basic definition of system as “a grouping of parts that operate together for a common purpose” ([25] p. 1-1). Moreover, he suggests that the way to organize knowledge is with this idea of system which are represented by the models we develop:

A structure (theory) is essential if we are to effectively interrelate and interpret our observations in any field of knowledge ... Without an organizing structure, knowledge is a mere collection of observations, practices and conflicting incidents ... A model is a substitute for an object or system ... Any set of rules and relationships that describe something is a model of that thing. In this sense, all of our thinking depends on models. Our mental processes use concepts which we manipulate into new arrangements. These concepts are not, in fact, the real system that they represent. The mental concepts are abstractions based on our experience. This experience has been filtered and modified by our individual perception and organization processes to produce our mental models that represent the world around us ([25] pp. 1-2, 3-1).

The previous statement summarizes core assumptions for SD. It points at the central notion of *mental model* and it emphasizes that these models are models of *systems*. In fact, in an article published in 1968 in *Management*



*Science*, Forrester [28] underlined that this application of feedback concepts to social systems was evolving toward a theory of structure in systems with a particular goal of policy design. Specifically, “industrial dynamics is a philosophy of structure in systems. It is also gradually becoming a body of principles that relate structure to behavior” ([28] p. 141). Two fundamental variables are identified: levels and rates, and the basic structural element is the feedback loop: “every decision is responsive to the existing condition of the system and influences that condition” ([28] p. 143). And, as stated above, the structure is an aid to organize knowledge in a particular situation given a particular purpose which is motivated by the pursue of explanation and policy design.

This historical review has accentuated central aspects of the foundation of SD. As a summary, this science initially envisaged by Forrester for designing systems is known nowadays as system dynamics. The models developed in SD have distinctive characteristics: dynamic structures, information flows, study of decision criteria, non-linearity, difference equations, symbolism and correspondence, and emphasis on confidence based on the structure of the model. The practice of this science is anchored on: the concept of servomechanism, the study of decision-making processes, the embrace of an experimental approach, and the use of computers for building formal models. These models are models of systems and the main goal is to help to organize our knowledge – which can be seen as arranged in mental models – so as to enhance learning processes and systems design on concrete settings and under specific purposes.

With these elements in mind, it is now possible to proceed with a brief discussion on important aspects regarding assumptions on reality and knowledge that can be recognized behind these premises and the practice of SD.

### “Real” World and Presentationalism

The traditional distinctions of *ontology*, *epistemology*, and *methodology*, form the habitual framework to drive a philosophical discussion. But the isolation of these issues may not be the most adequate or clear strategy. Given the interrelated nature of such categories and the misleading discussion on those terms which is present in a large part of organization and management science literature – part of this confusion will be exposed and clarified below – then a different plan will be used to develop the rest of this article. An instinctive option is to pick up significant issues and to relate them with what we can identify as part of the core of SD.

Where to start? Assumptions concerning a “real” world can be a first step to take since the traditional debate developed through the years around the claim of building models of “social systems” has fueled part of the discussions in systems science. As most of examinations on philosophical matters the debate has been permeated by a confusion originated in terms and words without the proper examination of the topic. However, this short revision is useful for opening the assessment of the premises of SD regarding a real world.

A prominent example is the criticism made by influential commentators who state that SD models represent an assumed “objective” real world [49]. This kind of critique usually labels SD as a “hard” approach, e.g. [48], meaning with such a term models of an assumed objective machine-like world – and habitually including and inverting the meaning of the term “positivism” – a mistaken assessment that still can be seen nowadays, e.g. [16]. This type of comments can be illustrated with the following quote from the work of Flood and Jackson [23]:

System dynamics models still center on capturing the structure of the “real world” ... the underlying assumption of SD that there is an external world made up of systems the structure of which can be grasped using models built upon feedback processes ... Because intentions derive from inside social systems, from the conscious human actors which constitute them, many possible appreciations of the nature and purpose of particular social systems are possible. SD simply does not deal with the innate subjectivity of human beings ... In essence the argument is that social systems cannot be studied, in the way of system dynamics, objectively from the outside (pp. 78, 79–80).

However, that is not what SD looks for. The mentioned emphasis on the examination of human decision-making processes and on the assumptions behind, the notion of mental model, the fact that model significance rests on its suitability for a particular purpose, among other aspects, should suffice to illustrate the point. Already Forrester in 1961 [24] emphasized: “a model can be useful if it represents only what we *believe* to be the nature of the system under study ... we are forced to commit ourselves on what we believe is the relative importance of various factors. We shall discover inconsistencies in our basic assumptions ... Thorough any of these we learn” (pp. 57–58, emphasis original). This should be enough to discard the assessment made by Flood and Jackson, and similar critics, who mistakenly placed SD in the terrain of a sort of naive realism as depicted in the quote above; this point is extensively clar-

ified by Lane [55,56]. More importantly, this argument is helpful to introduce the discussion about the nature of SD models, the assumptions behind these models, and their relation with a “reality”. The notion that drives these matters has been labeled in SD literature as “mental model” and even though this concept is not free of debate [21] it is placed at the core of the discipline.

The idea of mental model was already addressed by Forrester, as it has been indicated; he demarcated it in 1970 as the mental image of selected concepts and relationships of the world around us that we carry in our heads [26]; furthermore, “the mental model is fuzzy. It is incomplete. It is imprecisely stated ... [It] changes with time” (p. 213). Doyle and Ford [22], looking for a consensus on this subject, propose as a definition: “a relatively enduring and accessible [conscious], but limited, internal conceptual representation of an external system (historical, existing or projected) whose structure is analogous to the perceived structure of that system” (p. 411). Sterman [98] summarizes what the expression “mental model” refers to: “our beliefs about the networks of causes and effects that describe how a system operates, along with the boundary of the model (which variables are included and which are excluded) and the time horizon we consider relevant ... Most of us do not appreciate the ubiquity and invisibility of mental models, instead believing naively that our senses reveal the world as it is. On the contrary, our world is actively constructed (modeled) by our senses and brain” (p. 16–17). It should be noticed that these mental models may refer as well to planned or desired systems existing in the mind of the modeler [54]. The ultimate goal of system dynamics is to enhance our learning processes by testing and improving our mental models in a way that becomes consistent with the complexity of the systems that we face and design everyday [98].

Then, what is a SD model as related to some “real” world? Or how is it related to the notion of mental models? The usual option to answer these questions is to frame the discussion in the debate realism/anti-realism. This examination is even more relevant considering that some criticism labels SD as anchored on “realism”. A brief clarification follows.

### Realism? Anti-Realism?

Typical of latest debates in history of science concerns the dispute between the so-called *realism*, i. e. theories are true or false as descriptions of the world, and *instrumentalism*, i. e. theories are more or less adequate. This latter position is closer to pragmatism: theories are just instruments to systematize – and for many philosophers predict – ob-

servations, but theories are not claims about the world; or they can be also subjected to linguistic frameworks but still always truth-value-less. An overview of this debate is made by Leplin [60]. Yet, here it should be clarified that both positions and those definitions – habitually taken by historians of science – are two sides of the same coin. In short, both are forms of idealism. This assessment will be commented next.

On the one hand, scientific “realism” is usually defended using the method of abduction as source of knowledge, i. e. inference to the best explanation, which is just a form of induction and hence it is confirmationist, ultimately relativistic. It is not realism at all; it is exactly the opposite, i. e. idealism. For instance, take the influential ideas of Sellars [91] who defends an illustrative traditional position of scientific realism holding the source of knowledge on observation and relying on justification by induction for building theoretical frameworks; he emphasizes: “Laws in question are stipulated to be inductively established in the observation framework” (p. 313), the theories are then refined by empirical generalizations and what he calls further “injections” of images of the theory into the observational framework, in a typical process of instructional correction. A further refined and formal model of such a framework (proposed by Friedmann, holding a model-submodel relation) is commented by Morrison [66] who still holds, nevertheless, the search for truth and justification as support for his criticism and the necessity of confirmation [67]; see also for example the paper of Kukla [52] with a criticism to the ideas of Friedmann, yet also supported also on confirmation. In short, this “realism” is just idealism as we know it. Similar “realist” positions abound on the literature of the history of science. E.g. Smith [95] postulates a realist stance based on “common sense” but, since for him “science begins with observations” (p. 53), his realism ends up, indeed, trapped in sensations, i. e. idealism. Quantum mechanics does not escape the debate, e. g. arguing in favor of realism within the subjective theory of probability (the Copenhagen interpretation, Bohr, Heisenberg, etc.), Dickson [20] stands on the verification criterion for discussing on what he calls “quantum realism”; naturally such base can not succeed, which is anyway best self-explanatorily reflected in the first part of the title of his paper “An Empirical Reply to Empiricism”; simply there is no such reply. Another case is the criticism of Brown [13] to the deterministic notion of “realism” that Chermiak discards (who supports it on computer simulation of finite agents); the criticism of Brown shows a Galilean notion of realism, i. e. achievable true descriptions of the world via laws, but in a very “complex world” and thus, for him, inaccessible to agents with lim-

ited cognitive capacity; this Galilean view is found in large part of complexity science. In other attempts, for instance Schlagel [90] defends “contextualistic realism” which ends up in relativism: elements of the world have a conditional status relative to contexts and conditions, the existents are real relative to the particular structures and contexts on which they depend; his proposal is just idealism and can be better described as a sort of multi-phenomenalism, indeed relativism. The diverse “realist” positions *actually rooted on non-negotiable empiricism* seem endless; e.g. further examples are found in [1,59,75], and in most papers with the expression “scientific realism” in the abstract.

On the other hand, let us consider a supposedly opposite position, for instance instrumentalism. Instrumentalism actually ends up in relativism as well, e.g. with respect to the points of view where instruments can be applied; at the end the notions of “applicability” or “adequacy” become reference frameworks for establishing partial truths. Newton-Smith [68] proposes a conciliatory position, he calls it “modest realism” which ends up indeed in a sort of “moderate” relativism; a catalog of various scientific realisms can be found in that paper too, all of them addressing still the question of how such truths about a real world can be sustained (justified). In short, instrumentalists declare the dependence on observations, i.e. idealism, and thus this position is in fact sharing assumptions with the alleged “realists”. Leplin [60] summarizes: “some theory can be reduced to observation by defining or translating theoretical terms into terms that describe observable conditions. The remainder must be construed instrumentally” (p. 394).

The “realism” vs. instrumentalism debate is futile and yet it is perhaps one of the core discussions in philosophy of science. But unlike these influential historians of science, the presented dispute tends to be dismissed by several professional philosophers who see it as self-serving and unsophisticated; Fuller [33] underlines this assessment though he also illustrates the consequences of leaving such pointlessness discussions to endure. In this case, two seemingly unaware idealist factions argue about the best way to establish positive knowledge, e.g. either with a supposedly “true” description of the world (more precisely: *phenomena*, subjectivism) or by pragmatism (and again: *phenomena*, relativism). This shared *idealism* is the matter of the next section.

### Presentationalism

A broader debate can be assessed framed in the opposition between *realism* and *idealism*, see e.g. [72] and [77]. It will be shown that SD rests on the broad stance known as ide-

alism, i.e. presentationalism – this latter term is preferred here just for clarity [10].

Firstly, it should be commented the widely inverted and misleading use of both terms. And there is good company. Blackmore [10] shows various celebrities that became misusers of the expressions in question such as the former president of the American History of Science Society and prominent Harvard University professor, Erwin Hiebert, and Sir Russell Brain, outstanding neurologist and former president of the British Association for the Advancement of Science. Add the seemingly customary tendency to quote references in second hand without inspecting direct sources and a few decades later we have the terms used in exactly their opposite original sense in journals and books. Blackmore pictures the situation:

Like-minded ‘empiricists’ have restricted what they understand by the term to what idealistic philosophers of science *have wanted them to understand by it*. And since the term ‘realism’ sounds good to ‘tough-minded empiricists’ and since idealistic philosophers such as Hume, Comte, Schuppe, and Mach and their recent successors *scarcely if ever have admitted their idealism*, many scientists and historians of science have let themselves be seduced into reversing the normal epistemological definitions of ‘realism’ and ‘idealism’. Even worse, respected scholars such as Stillman Drake, perhaps our most outstanding authority on the manuscripts of Galileo, and Larry Laudan, a young, energetic, and much published commentator on Mach and ‘empiricism’, have allowed themselves to become advocates of hopelessly naive ‘non-philosophical’ positions. Drake is sure that Galileo held no philosophical position, or that if he did, it had no effect on his scientific work. Laudan is equally positive on the basis of Mach’s written comments that his phenomenalist epistemology had no influence on Mach’s ‘empirical’ methodology of science. The simplicity of their views can partly be explained by their tendency to understand by the term ‘philosophy’, not one’s most basic universal assumptions, *but expressed talk about speculative matters*. Similarly, many ‘materialists’ who feel sure that the physical world is directly given in experience, and who accept the idealist Kant’s distinction between ‘science’ and ‘metaphysics’, are convinced that anyone who identifies the physical world with what is *beyond* immediate experience is an ‘idealist’ and ‘metaphysician’ and (following Wittgenstein) ‘is merely uttering nonsense’ (p. 131 in [10]).

In order to clarify, it is appropriate to underscore the attitude behind an empiricist epistemology. A major defining posture is what has been labeled as *idealism*, given the natural disbelief of a world beyond senses, which is the pillar of an empiricist epistemology. The term “idealism” comes from the “idea” of Bishop Berkeley, who took physical objects as “ideas” which included sensations and thoughts: “It is evident to any one who takes a survey of the objects of human knowledge, that they are either ideas actually imprinted on the senses, or else such as are perceived by attending to the passions and operations of the mind, or lastly ideas formed by help of memory and imagination, either compounding, dividing, or barely representing those originally perceived in the aforesaid ways.” [8]. This idealism relies entirely on senses and mind-dependent worlds since consequently sense-data were the only things of whose existence our perceptions could assure us, and that to be known is to be ‘in’ a mind, and therefore to be mental. Berkeley, therefore, concluded that nothing can ever be known except what is in some mind, and that whatever is known without being in my mind must be in some other mind [85]. Hunter [47] summarizes:

As a result of their constraints on knowledge and meaning, empiricists tend to be skeptical of necessary truths that are independent of mind and language, and of putative eternal abstract entities (p. 110).

This idealism can be equally identified with terms such as ‘phenomenalism’, ‘neutral monism’, or ‘subjective idealism’ (e.g. [10]), or presentationalism. In other words, “anything in time or space, anything than can be known by the human mind, is phenomenal” (p. 146 in [15]).

Taking this posture to the context of science, Bartley [4] provides the implications: “Presentationalists see the subject matter of science not as an external reality independent of sensation. The subject matter of science is our sensory perceptions. The collectivity of these sensations is renamed ‘nature’ ... The aim of science is seen not as the description and explanation of that independent external reality but as the efficient computation of perceptions ... [It] became the dominant twentieth-century *philosophy* of physics” (p. 11, 16, emphasis original). In general this position is the pillar of the prevalent conception of science which has been fueled by physics (for instance in the interpretation of quantum mechanics of Bohr and Heisenberg) and backed by influential names like Mach, Russell, Wittgenstein, Ayer, Lewis, Carnap, etc.

To appreciate the contrast, a standard definition of *realism* can be:

‘Realism’ ... is used for the view that material objects exist externally to us and independently of our sense experience. Realism is thus opposed to idealism, which holds that no such material objects or external realities exist apart from our knowledge or consciousness of them, the whole universe thus being dependent on the mind or in some sense mental. It also clashes with phenomenalism, which, while avoiding much idealist metaphysics, would deny that material objects exist except as groups or sequences of sensa, actual and possible (p. 126 in [10]).

Apart from the particular emphasis on material objects (as opposed to Berkeley’s *ideas*), another point is that realism defends a cosmocentric thesis opposed to the anthropocentric view in the discussions of the alleged “realists” of science presented earlier; in the latter case the observer-centered learning process is fundamental, and it is carried on via induction looking for acquiring positive, verifiable, and true knowledge – or justified true belief. Moreover, let us recall that historians of science denote with the term “realism” just the concern with supposed true descriptions of the world. But in fact *realism* does not imply that knowledge is achievable, it does not imply that the world is a perfect clock, it does not imply determinism, it does not imply that there can be correspondence between theories and such real world; these are different affirmations that unfortunately seem to be muddled inside the same bag. One thing is to assume a real world beyond senses. But a different inquiry is the character we ascribe to it. Another different issue is the role we assume for our senses. Another very different concern is the question of knowability, etc. Rescher [77] illustrates typical examples of the confusing use of terms in literature: “The three positions to the effect that real things just exactly are things as *philosophy* or as *science* or as ‘commonsense’ takes them to be – positions generally designated as *scholastic*, *scientific* and *naive* realism, respectively – are in fact versions of epistemic idealism exactly because they see reals as inherently knowable and do not contemplate mind-transcendence for the real” (p. 187).

Coming back to presentationalism, this position then assumes that we are imprinted by the environment, and we call to this *impressions* “knowledge”. Given the limitation of our senses then presentationalism postulates that nothing more can be known; and thus, such assumption is used to construct the world in our minds: for a presentationalist, strictly speaking, the world is not re-presented since we do not have access to it, the world is just what is *presented* to our senses: the world as we experience it happens to be the world itself; and, since anything that can be known by

the human mind is, then, phenomenal (sensations, etc.), therefore knowledge strictly depends on – and is source in – what is sensed, e. g. observed. Hence knowledge needs to be justified in order to avoid error; and yet, the only existent knowledge is the imperfect evidence sourced in sensation and, nevertheless, a foundation that can be justified is pursued. This is the popular plan we have come up with, so far, to try to avoid the destruction of empiricism made by Hume. The picture can be summarized:

Almost all traditional epistemologies are Lamarckian in their accounts of the growth of knowledge. This is conspicuously true of presentationalism, almost all adherents to which maintain an inductivist, justificationalist account of knowledge growth, according to which knowledge is constructed out of sensations (as building blocks or elements) by a relative passive process of combination, accumulation, repetition, and induction (p. 25 in [4]).

This position is identified with *idealism* and most popularly associated with *epistemological empiricism*, with all its assumptions and its consequent scientific method.

The last point to underline is that this epistemology has subordinated ontology. A remarkable illustration of this type of problems was already made by Bowman [11]: “The result ... is the more or less deliberate abnegation of a genuine epistemology and the substitution for it of a highly formal logic. Hence the paradox illustrated equally in the case of Plato and, recently, of Mr Russell, of a radical empiricism (expressed in Plato’s Protagorean theory of sensation and in Russell’s subjectivism) subsisting side by side with the extremist rationalism. Such a dualistic position is the despair of philosophy, and indicates a failure in the synthetic work of thought” (p. 485). This despair is easy to recognize in current science which presents a contradictory ontological position whose difficulty is found in its subsumption under a radical epistemology. Indeed extreme empiricism, i. e. presentationalism, has become the metaphysics, i. e. the theory of reality, of our science. The debates on “scientific realism” presented earlier picture this failure. On this particular subject the different use of terms in literature is a source of confusion; but here the terms have been inverted by historians of science and scientists; and beyond a semantic confusion this has brought a narrow conception and a very restricted examination of epistemological assumptions. In short, the subjectivism of Descartes and Kant – or more precisely, Kantian idealism – is what now is labeled as “realism”, e. g. *everything* has become “phenomena”. As a matter of fact common expressions like “objective phenomena” or “real phenomena” uncover an idealistic position where

the “objective” or “real” are just *phenomena*. Indeed, the so-called scientific “realism” commented above is nothing more than *a sort of empiricism* driven by Hume and Kant. This “realism” is just ontology overlapped by epistemology (*idealism*). More precisely, regarding Kant, let us recall his “Transcendental Idealism” which was the common answer of Kant when he was accused of idealist, e. g. see [93], denying to be a “dogmatic idealist” (in the Berkeley sense; see e. g. [101]); a full discussion of this failed defence of Kant is made by Guyer [37]. In particular Turbayne [101] defends Kant when he was accused of having misinterpreted (or even completely having misunderstood) Berkeley’s idealism; yet, Turbayne’s conclusion summarizes the Kantian idealism (and its ambiguity) unmistakably: “The Kantian antidote to this is not the a priori nature of space, but *its reality or subjectivity*, which assimilates space and its contents into the realm of ideas, and thus *prevents illusion*” (p. 243, emphases added). Regarding consequences, perhaps the best summary is the radical idealist position of Mach – who denied even the existence of atoms since they cannot be observed. Kant seems to be taken for granted without a proper reflection on his position.

These few points were commented since this debate is the major informer of the method and the assumptions of management science, organization science, and social science in general, places where SD has its roots. More important, by making this clarifications then a clear ground for SD can be envisaged. It can be affirmed that SD has been mistakenly labeled as “realist” by many commentators alluding the alleged aim of building “true” descriptive theories of the world, and using the misleading definitions of historians of science. But also from the discussion above it should be clear that SD safely rests on presentationalism. Moreover, the identified presentationalist stance known as “instrumentalism”, and in general the so-called anti-realists positions (independent of the inverted use of the term), fit to the SD worldview: the abstractions from our experience are arranged in mental models which form knowledge that we want to improve in order to make better decisions. The SD models built for achieving this goal are judged against their adequacy and suitability for a particular purpose; these models are not claims about the world but instruments for systematizing observations and for boosting learning processes using experimentation via simulation.

### The Discussions on Positivism: Presentationalism and Knowledge

An issue previously mentioned which is closely connected with presentationalism is relativism and positivism. Since



positivism usually – and mistakenly – is pejoratively associated with a supposed objective representation of reality, then an important clarification is needed. In short: positivism is consistent with presentationalism and with relativism as well.

### Presentationalism Brings Positivism

Blackmore [10] reminds that strictly speaking neither “rationalism” nor “empiricism” are properly epistemological terms at all; the entrenched idealism has led just to this narrowed identification. For instance, usually the term “empiricism” is synonymous of knowledge sourced in observation, i. e. in a restricted epistemological context, but such popular narrowness is inaccurate. Blackmore remarks: “Granted, that if one *means* by ‘empiricism’ not just an extensive and careful concern with empirical evidence but *restricting* reference or knowledge or both to sensory appearances, then there are indeed epistemological implications. One has become an epistemological phenomenalist or subjective idealist, or if you will, a positivist” (p. 130, *emphases original*). This clarification is needed for two reasons; on the one hand, as it was stated, SD has been labeled as “positivist” but the critics take this term as a sort of naive realism; the confusion is patent once we realize that positivism actually is a consequence of idealism, the opposite doctrine of realism. On the other hand, since SD is better identified with idealism than a sort of positivism can be also associated with it, but not the sort of “positivism” that the critics have in mind but the authentic positivism; and yet we will see that with the use of simulation positivism does not necessarily fit either.

The case of management science is a good example regarding the discussion on relativism and positivism. Let us consider the traditional and unfortunate sharp division between “hard vs. soft” which also takes the form “quantitative vs. qualitative”. However, this discussion is misleading as well. It is not difficult to find researchers that claim to be anti-positivists but being themselves grounded on positivism (e. g. empirical observation, verification, induction, etc.) without noticing the contradiction. A good example is the claimed opposition between positivism and phenomenology. Yet, phenomenology is authentic positivism when is committed to evidence – Husserl himself underlines this aspect – see e. g. [94]. Indeed it is easy to appreciate an inverted use of the term “positivism” in literature; in management science this is a favored and widespread practice where so-called anti-positivists do not notice their positivism. The fact is that anthropocentrism is the ground in our most influential epistemologies that recognize the obvious imperfections of our sensorial apparatus but, nev-

ertheless, rely knowledge on sensation (observation, etc.), that is, positivism, which is nothing more than our anxiety to confirm and validate, i. e. justify, our “imperfect” knowledge, e. g. empirically. This is a simplification of highly loaded terms; clarifications and further discussions can be found elsewhere, e. g. regarding positivism see [9,97,100].

### Positivism is Anchored on Justification

Within a presentationalist worldview the search for confirmation and verification is nothing less than the search for justification of knowledge where the intellectual authority lies in sense experience. From a presentationalist account it is straightforward to have a justificationist approach for confirming and verifying theories. Following Bartley [4]:

Preoccupied with the avoidance of error, they suppose that, in order to avoid error, they must make no utterances that cannot be justified by – i. e., derived from – the evidence available. Yet sense perception seems to be the only available evidence ... The claim that there is an external world *in addition* to the evidence is a claim going *beyond* the evidence. Hence, claims about such realms are unjustifiable. Crucial to the presentationalist argument are, then, two things: the desire to give a firm foundation or justification to the tenets of science, and the construal of sense experience as the incorrigible source of all knowledge (pp. 12–13, *emphases original*).

In fact justification philosophy taken as the search for epistemic ‘authorities’ has been the dominant style of western philosophy looking for “well-grounded” knowledge. For instance in the customary view of knowledge as *justified true belief*, e. g. in the sense of Russell [86]) – as the result of systematic analysis “of our sensory experience of a knowable external reality” (p. 47 in [96]). Within this popular position the central problem of epistemology – as succinctly formulated by Radnitzky [76] – becomes:

“When is it rational or, so to speak consistent with one’s intellectual integrity, to *accept* a particular position?” The formulation suggests the direction in which the answer is to be sought: “When concerned with a statement, a theory, etc., accept those and only those statements, theories, etc., which not only are true but whose truth has been established” (p. 282, *emphasis original*).

The goal of justification is usually entrenched within the method of induction where every new repeated observation is a confirmation that validates – justifies – the theoretical statement. Even with weaker conditions the way

of reasoning is the same, for instance within the ideas of Ayer where strict verifiability is seen as a too rigid criterion – he introduces *confirmability to some degree*, instead of complete and conclusive verifiability (see [88]); justification is still pursued. The appeal of justification can be explained because it looks for avoiding relativism (inherently attached to presentationalism) since not all positions are equally good or bad and it suggests to look for something beyond blind belief [76]. Though it is not the only option, nevertheless, it is the most common view of science; the concerns on validation, justification, verification, confirmation, and generalization, are part of this popular and influential view. Here the observer is the fixed point of reference. In short, within a justificationist logic, it is rational to accept only those positions that have been justified according to the rational authority which in the case of presentationalism is sense experience, consistent then with the highly influential ideas of Locke, Berkeley, Hume, Mach, Carnap that have shaped our prevalent view of science [4].

### Justification in System Dynamics

Turning back to SD, it must be recalled the role of mental models whose characteristic nature of “abstractions based on experience” can be better assessed within a presentationalist stance. Here justification has also a place. How is this knowledge justified? Already Forrester [24] emphasized, within the debate of model validation, that, “knowledge of all forms can be brought to bear on forming an opinion of whether or not a model is suitable to its particular purpose” (p. 129). Therefore, Forrester [25] also emphasized that “we can never prove that any model is an exact representation of ‘reality’ ... Models are then to be judged, not on an absolute scale that condemns them for failure to be perfect, but on a relative scale that approves them if they succeed in clarifying our knowledge” (p. 3–4). This sort of relativism will be addressed next.

With the aim of placing this stance within the discussion of history of science, Barlas and Carpenter [3] addressed the “philosophical roots of model validation” associating SD with what they called a “relativist philosophy of science”. In this view justification is pursued: such a knowledge is seen as socially, culturally and historically dependent and it becomes socially justified belief. Here “a valid model is assumed to be only one of many possible ways of describing a real situation ... for every model carries in it the modeler’s world view ... validation is a matter of social conversation” (p. 157). Hence, confirmation and verification are pursued through a social process relative to a frame of reference. This sort of moderate rela-

tivism was later criticized by Vásquez, Liz and Aracil [103] for whom such relativism is unacceptable in spite of their recognition that there is no privileged single model (or set of models); since they are also concerned with epistemological justification these authors present Putnam’s internal “realism” as a more adequate way to conceptualize the type of knowledge consistent with the assumptions about reality held by SD practitioners; in short, these authors underline that mental models are the source of knowledge – and its justification – helping to select the structures that must be assumed as working in real systems; here knowledge is taken as internal to the conceptual scheme of SD. Since there can be many models for a given situation, the authors argue that this framework gives the possibility of convergence as a result of “the strong interactive character of mental models” (p. 34) recognizing that in any case SD modeling is a process of revision and adjustment. With this proposal these authors seek to achieve justification and some realistic representational content in spite of the plurality of alternative SD models available for a specific situation. It is easy to see that this proposal is still relativistic, in this case knowledge is relative to the mental models and the conceptual scheme – though the mentioned authors would not agree since for them there is a “reality” given by the internal representational schema, in this case the mental models of the modelers.

The fail to recognize presentationalism as the epistemology which is driving ontology is at the root of the presented discussions. This is a distinctive trait present in large part of the philosophy of science literature. However, the characteristic problem of mistaking positivism with a sort of supposed “objectivism” is also present in this discussion – in fact, Barlas and Carpenter, following the misdirecting literature on the subject, argue that the relativist philosophy that they defend rejects positivism; Vásquez, Liz and Aracil also follow the same inertia. However, these theses, which seek for confirmation, verification, justification, reflect the search for *positive* knowledge within an idealist epistemology. In the first case, knowledge is confirmed and accepted through social interaction and it is relative to a context. In the second case, knowledge is justified on mental models. To appreciate a genuine contrasting position see an anti-justificationist approach to validation in computer simulation in [51]; the core of anti-justificationism can be found in the work of Popper, e. g. [73,74].

### System Dynamics as Presentationalist

It should be clear that the assumptions on SD reject naive realism; the models are not supposed to be accurate and

corresponding descriptions of an external true reality – and furthermore, this is not what the term “positivism” refers to. In any case validation in SD does not mean a supposed “positive proof” or to assess the development of “true models” of the world. On the contrary, SD aims at enhancing our ways of reasoning, it emphasizes a process of learning so as to consistently improve our mental models which are product of our experience and the operations of our mind. Hence SD is closer to presentationalism. Knowledge can be socially justified and our mental models can be enhanced. Moreover, a particular emphasis on the modeling *process* is also underlined – see e.g. [29,44,89] and it will be commented in the eight section devoted to simulation.

### A Brief Note Regarding Social Theory

Another important discussion is related to the theories about the “social world” that are supposedly held in SD, i.e. the social theory behind SD if any. Part of the misguided debate is explained by the widespread use of the traditional framework of Burrell and Morgan [14] whose oversimplification of social science in four paradigms has deviated major issues covering important topics under a too practical and inadequate schema – e.g. see a criticism in Deetz [19]. Jackson [49] provides a summary of such usual misconception:

System dynamics ... is essentially functionalist in nature. It sees system structure as the determining force behind system behavior ... If humans are free to construct social systems as they wish, what determining influence does system structure have? ... This tension between determinism and free will is unresolved (p. 39).

It should suffice to recall from the discussion above that the *aggregate* approach of SD is not a theory of human behavior; SD is not concerned with *individual* action. Furthermore, it does not assume that a structure, of any kind, determines human behavior either, i.e. the sort of determinism that Burrell and Morgan [14] oppose to “free will” and in the lines of the already vague term known as “structuralism” – see an early clarification of the problems of such type of oversimplification in [84]. In any case, this sort of criticism has been answered and clarified by Lane [56] who has underlined the main point: SD is concerned with aggregate social phenomena and not with individual meaningful actions [55]. Moreover, system dynamics does not propose invariant causal laws, as Lane [56] also concludes: “The only universal law/theory on offer is a grand methodological, or structural theory,

associated with a representation scheme ... it does not attract the determinism-related criticisms attached to grand theory in the sense of Parsons and Mills” (p. 111). In [57] Lane proposes to link system dynamics with a different framework: agency/structure theories.

### Servomechanism

Part of the confusion is because of the misunderstanding by various commentators of the notion of feedback that underlies SD. This point has been also a source of misconception given the use of feedback in theories of control applied to social systems, e.g. cybernetics. Consider for example the following comment by Flood and Jackson [23]:

The attempt of SD thinkers to model external reality is misguided ... The emphasis placed on “structure” as the means of revealing knowledge about the optimal behavior of systems cannot be accepted ... SD modelers using feedforward control appear to believe that there are optimal future states that we should steer systems towards” (pp. 80, 81).

Such criticism apparently is associated with the notion of feedback used in cybernetics. However SD does not pursue optimization, let alone by studying “knowledge revealing” structures in order to achieve supposed optimal behavior patterns. Instead, the goal is to have a better understanding of feedback structures in order to enhance decision-making and policy design. This point has been also addressed by Lane [53,55]. The central clarification of this issue has been made by Richardson [78] who distinguishes two different threads in the development of the concept of feedback in the social sciences: the cybernetics thread and the servomechanisms thread. The failure in noticing these two different lines of thought has produced various misconceptions regarding the notion of feedback in SD, a concept that has been shown as one of its building blocks. On the one hand, the cybernetic conception of feedback is defined in terms of input and output, it is limited only to loops of negative polarity which in turn are conceived as the mechanisms of control – and hence there is a particular interest in goal seeking and goal formulation given the concern in cybernetics for achieving adaptive behavior via directed processes and homeostatic mechanisms; feedback mechanisms guide this pursue of viable behavior which is carried by goal-seeking processes. On the other hand, in SD, coming from the servomechanisms thread, feedback loops are taken as intrinsic parts of the system (and not just as mechanisms of control), it includes also loops of positive polarity, and such feedback structures are seen as responsible for counterintuitive behaviors and policy resistance in

social systems; here the analysis is directed toward policy design.

### Explanation and Mechanism

The next interesting question would be how we can achieve better understanding and better policy design by enhancing our mental models. How can we characterize this type of knowledge?

A solid account of *explanation* should be placed at the heart of any scientific activity. The general inquiry about [scientific] explanation has to do with “learning how the process of doing science facilitates understanding, and what type(s) of understanding science provides” ([7] p. 307). In a very intuitive way, a first approach to explanation might be associated plainly with removing puzzlement [6]. It is also common to affirm that an explanation aims to answer queries of *why* in order to provide understanding [87]. Yet, to characterize such idea is a major and open unresolved question in philosophy of science [69]. The notion of causality has traditionally played a central role; this view has pervaded most of scientific research where theory development and explanations are essentially conceived as the search for causes, e.g. [50,87]. However, several explanations are not essentially based on simple causal relations but on other approaches such as identification, models, analogies, formal linguistics, laws of association, laws of co-existence, variational principles, among others [83]. Berger [7] underlines the prominence of this question when attempting to characterize the explanations provided by nonlinear dynamical modeling:

Mathematical modeling [is recognized] as one of the central activities of science, and it is reasonable to say that modeling explanations dramatically increase our understanding of the world. But the modeling explanations found in contemporary scientific research show that the interesting claims of causal accounts are untenable ... An adequate account of scientific explanation must accommodate modeling explanations, because they are simply too central to ignore (pp. 329–330).

The main goal of this section is to explore the position and the characterization of the kind of explanation pursued in system dynamics. This characterization fits with the presented core of SD and the presentationalist stance. And again various clarifications will be needed along the way.

### Causality

The difficult issue of causality can be treated in several senses. In the first place, a possible relationship associated

with the term “determinism” on human behavior is dismissed with a previous argument: SD causal models does not point at supposed laws of causality governing human action [55,56]. What is more interesting is to investigate the concept of causality as such in SD models; after all, a large part of SD modeling relies on what are known as “causal”-loop diagrams. Forrester emphasized the term *interrelationships* [24] where feedback loops are understood as closed informational paths connecting in a sequence decisions that control actions [25]; he labeled it as a “circular cause-and-effect structure” (p. 1–9). In fact the development of *causal*-loop diagrams has become important in SD practice; in particular, flow diagrams were initially recognized as useful pictorial representations that help to formulate and communicate the structure of a dynamic model [24]. These models are ultimately theories of behavior, surely causal theories of behavior. But consistent with presentationalism, these theories are sourced in the mental models of the modeler, there is no direct connection to an alleged causation in a real world. Sterman summarizes a definition: “a causal diagram consists of variables connected by arrows denoting the causal influences among the variables” [98]; here, every link represents *what the modeler believes is a causal relationship* between two variables; this causal attribution is seen as a central feature of mental models, as Sterman also stresses “we all create and update cognitive maps of causal connections among entities and actors ... Within a causal field, people use various cues to causality including temporal and spatial proximity of cause and effect, temporal precedence of causes, covariation, and similarity of cause and effect” [98]. Again, the core of the discussion should be driven by the concept of mental model in order to deliver a clear discussion on this view of causality held in SD. This section outlines a framework.

Most of the time we seem to hold a strong causal view of the world. In particular, the causal relationship – whatever that could be – tends to be the source of explanatory power, i.e. the *explanans*, and one usual source of validity, that is, *for having a relevant valid explanation we must have a causal relationship*. That is the usual principle, and it is usually associated to the term “determinism”. Recalling the discussion on presentationalism, one should note that – as Hesslow [42] underlines – if we are going to retain a Humean view of the world, i.e. consistent with idealism, then it seems that we have two different paths, probabilistic or deterministic. The latter one is of interest here (for the probability account of causation see e.g. [43]). Within a deterministic approach, a cause is always sufficient condition or a part of a sufficient condition for the effect, that is, if  $A_t$  is a nonsufficient cause of  $B_{t'}$ , then there must be



some auxiliary condition  $C_{t''}$ , [with  $t'' < t'$ ], such that  $A_t$  in combination with  $C_{t''}$  is sufficient for  $B_{t'}$ . Hesslow clarifies this “sufficiency principle” [42]:

The deterministic approach has been something of a received view of causation. This view, which we may call the ‘sufficiency principle’ is also common among scientists. The sufficiency principle is not in itself strictly deterministic. It does not mean that every event has a sufficient cause, only that if an event has a cause, then it has a sufficient one. However, it seems that the popularity of the sufficiency principle is a reflection of a widely spread, though usually implicit, commitment to the stronger thesis, that every event has a sufficient cause (p. 592).

The cited paper of Hesslow aims to show that the deterministic approach is superior to the probabilistic one, that is, the idea that the probabilistic account presupposes determinism. Indeed what is happening is the trap of Hume – so to speak. The issue in hands is illustrated with the *Humean fork*: based on observation of constant conjunction of events – altogether with temporal priority, i. e. the cause is observed prior to its effect – we *suppose* a causal connection between them – see summaries in e. g. [46,70] and the original work of Hume [45]. These *suppositions* are arranged in our mental models, that is, in our theories about what we assume as the relevant causal connections that we suppose so as to explain the world.

With this framework in mind, the network of causal connections that the SD modeler believes to be relevant indicates a sort of “sufficiency principle”, but of a special kind. It is sufficient relative to the purpose of the model, as it has been indicated above. And it is sourced in the mental model since SD is seen as a vehicle for learning and not as a device for operating a “real world”. How are these causal relationships portrayed? In a generic form, a feedback loop based stance is based on the fact that decisions in a time  $t$  affect the environment which affects again the future assessment of the situation in a time  $t'$  ( $t' > t$ ) which usually is the base for new upcoming decisions and actions taken in a time  $t''$  ( $t'' > t'$ ) and so on. These relationships are not necessarily close in space or time. Furthermore, a related issue is what can be labeled as “multiple causality”. The complexity of social systems is a current concern for social scientists; the multiple interactions among several agents, actors, or entities, is what has been distinguished as the key to study and to understand complex systems because of the recognition of our inability to deal with them based on traditional incomplete simple-causality thinking – the assumption that explanation of phenomena can be satisfactory or even sufficient based in simple unidirectional

causal relationships between variables or constructs. What is more, feedback loops have important implications associated with counterintuitive behavior that usually we do not consider easily or that we misunderstand; they have a key role in complex settings; in fact they are, to a large extent, responsible for the arising of complex behaviors; this is a central affirmation in SD [26,98]. The simplest example might be the tendency that we have to infer linear growth from single first order positive feedback loops. But the behavior here actually is exponential. Let  $S$  be the state of a system and  $g$  the constant fractional growth rate, the linear first order positive loop and its solution are:

$$\frac{dS}{dT} = gS$$

and it has as solution:

$$S = S_0 e^{gt}.$$

The central question is: are these causal theories, portrayed in system dynamics, claims about the world? i. e. Are these models assumed true descriptions of the world? Clearly no. From the presentationalist stance of SD, the causal-loop diagram is essentially what the modeler *believes* is the relevant causal network for the problem in hands. It constitutes his theory about it. The source of knowledge is the mental model and causality is only a supposition of the modeler and a way to arrange knowledge, it is not an affirmation of truth about a supposed causal world. And furthermore, causation as such is not the source of explanation provided by SD. In order to clarify this we should take a look to the notion of explanation held in system dynamics.

### Mechanism

System dynamics aims to answer *why* questions. This is done generally via the development of *dynamic hypotheses*. A core premise of SD has always been to enhance learning and to provide understanding [30]. This aim has been stated from the very beginning; for instance Forrester asserted in *Industrial Dynamics*: “Our objective is to enhance understanding and to clarify our thinking about the system” p. 57 in [24]).

How is this goal pursued? A SD model should be able to account for a specific problematic behavior which is explained in terms of the structure of the model; here the term “structure” refers to the stock and flow organization, the feedback loops and the rules of interaction [98]. This approach to explanation is known as a dynamic hypothesis and is the core concept in order to provide understanding from a system dynamics point of view; its endogenous



character is the chief feature that makes it intelligible; for instance: “One key task in this search for insightful, system level understanding is the telling of ‘system stories’ – coherent, dynamically correct explanations of how influential pieces of system structure give rise to important patterns of system behavior” (p. 1 in [65]). In fact this task represents one of the more significant research lines [79].

How can we characterize this particular kind of explanation? As a first point, the notion of organized social complexity helps to drive this discussion. A quote borrowed from Hayek illustrates it:

Where we have to deal with such social wholes we cannot, as we do in the natural sciences, start from the observation of a number of instances which we recognize spontaneously by their common sense attributes as instances of ‘societies’ or ‘economies’ ... What we group together as instances of the same collective or whole are different complexes of individual events, in themselves perhaps quite dissimilar, but believed by us to be related to each other in a similar manner: they are classifications or selections of certain elements of a complex picture on the basis of a theory about their coherence (p. 43 in [40]).

Based on the quote above of Hayek, he suggests conceiving the *explanation* as modeling [104], and for social sciences he rejects the usual prediction and control aspirations and asks the reader to focus more on models to explain typical processes [64], he depicts it with biology: “It deals with pattern-building forces, the knowledge of which is useful for creating conditions favorable to the production of certain kinds of results, while it will only in comparatively few cases be possible to control all the relevant circumstances” (as cited in p. 202 in [104]). Hayek calls it “explanation of the principle”. Essentially he means the explanation of a *kind* of phenomena instead of particular events. As another example consider mathematics: “A set of equations which shows merely the form of a system of relationships but does not give the values of the constants contained in it, is perhaps the best general illustration of an explanation merely of the principle on which any phenomenon is produced” (p. 291 in [39]). This analogy illustrates the notion of abstract relations that would build an “explanation of the principle” which can be associated with *mechanism* as the source of explanatory power – instead of causality.

Again, a clarification is needed given the widespread identification of the term “mechanism” with ontic commitments. Fundamentally mechanism is a kind of *explanation*. It should be noticed that “mechanism” refers to *epistemological* issues. However, the term is habitually associ-

ated with assumptions about reality. But Hogben already clarified in 1930:

In any discussion between the two [mechanist and holist or vitalist], the combatants are generally at cross purposes. The mechanist is primarily concerned with an epistemological issue. His critic has always an ontological axe to grind. The mechanist is concerned with how to proceed to a construction which will represent as much about the universe as human beings with their limited range of receptor organs can agree to accept. The vitalist or holist has an incorrigible urge to get behind the limitations of our receptor organs and discover what the universe is really like (1930, p. 100, as cited in [12], p. 347).

The explanatory notion of mechanism is well underlined by Grene: “Let us look for a mechanism which might underlie the phenomena we hope to understand, seeking wherever we may relevant sources from which to derive ... an analogue of a possible mechanism ... [Such an explanation is of value because it tells us] *how in fact those phenomena are produced*” (as cited in [12], p. 346, emphasis original).

However, there is still no agreement about what a mechanism is and how it appears to succeed in science as a way to provide understanding. Perhaps the most complete account is the one of Tabery [99] who proposes integrating two complementary points of view. These two aspects are (i) the interactions among several parts and, (ii) the activities associated with these interactions. Both characteristic are taken as necessary for having a mechanism-based explanation.

On the one hand there is the emphasis on interactions, a thesis supported by Glennan [36] with a central concern on the nature of *complex systems* since the role of parts and its interactions are conceived as indispensable. The work of Bechtel and Richardson [5] develops the association of mechanism and complex systems (in biology and psychology) and emphasizes the tasks of decomposition and localization as the heuristics in order to uncover mechanisms. This position replies to the conventional view of a mechanism as merely the interactions between causal processes as the essential *explanans*; Glennan [36] stresses: “A mechanism for a behavior is a complex system that produces that behavior by the interaction of a number of parts, where the interactions between parts can be characterized by direct, invariant, change-relating generalizations” (p. 344). Glennan clarifies also that he avoids using the term “causal-law” and instead uses “change-relating generalizations” because these relations are not exception-less as traditionally a law is understood. In addi-

tion, he emphasizes the very different character of such account which is in opposition to the traditional causal view in which mechanisms are sequences or chains of events leading up to a particular event – which is often associated in systems theory literature with “linear thinking”.

On the other hand we have activities. Machamer, Darden and Craver [62] emphasize this aspect adding that mechanisms are not only inter-connected entities but also activities producing regular changes from initial to finish conditions; they call themselves dualists since for them both notions – entities and activities – are necessary to constitute a mechanism: “The organization of these entities and activities determines the ways in which they produce the phenomenon. Entities often must be appropriately located, structured, and oriented, and the activities in which they engage must have a temporal order, rate, and duration” (p. 3). It is important to underline their critique to Glennan’s view arguing that the concept of *activity* is fundamental to understand the changes produced (because of the activities) through the process and not only as the black-box view of change of states or change of properties of the inter-connected entities; they picture it clearly with the following statement: “it is not the penicillin that causes the pneumonia to disappear, but what the penicillin does” (p. 6). Furthermore, in order to account for a mechanism they emphasize three distinctions: set-up conditions (as *part* of the mechanism, not as a sort of input; this includes relevant entities and their properties and initial states), intermediate activities (including also relevant entities, properties, and an intelligible account of the activities that link them) and termination conditions (such as privileged endpoints, equilibrium states or the final stage of some unitary integral process). They also draw attention to the fact that mechanisms take place in nested multi-level hierarchies and that they usually are not full pictures but truncated abstract accounts – a *mechanism schema* – depending on the required level of detail or aggregation.

System dynamics modeling is perhaps one of the best ways to picture this kind of explanation. One can distinguish two main components in the structure of SD models: the physical and institutional assumptions – including the chosen parts/variables and the interconnections between them, and the decision rules of the agents [98]. The interplay between the physical structure and the associated decision rules as the explanation for behavior is a foundational aspect. Indeed the interconnections and the activities needed to account for a mechanism are included in these system dynamic models structures. Specifically, the activities producing change can be referenced in the links of the models and the decision rules describe how the in-

teractions produce certain activities. The whole set of initial and final conditions and the inter-connected parts engaged in producing activities characterize a mechanism. For example, the simplest mechanism is perhaps a single feedback loop. The set-up conditions are the initial values of the variables involved, the termination condition is the endpoint of the loop which can be accounted in a mechanism as “the final stage of what is identified as a unitary, integral process” (p. 12 in [62]). The intermediate activities are depicted by the links and the application of the decision rules. For instance, a simple positive first order loop can produce exponential behavior beyond the particular values of the variables involved. More intricate structures are the source of different behaviors, i. e. the change in the values or patterns of variables through time.

This stance fits an explanation of an abstract principle in the sense of Hayek. The structure is the source of explanation of patterns of behavior, i. e. the change in the values or patterns of variables of interest through time. This is known in SD as a dynamic hypothesis [98]. With this focus on aggregate patterns – instead of individual events – as consequence of the structure, it can be said that the “causal” mechanism is indeed the loop structure of the system, or the particular and relevant feedback substructures of the model that may explain the behavior; for instance Richardson [78] illustrates it in this way: “The ‘cause’ of an arms race is viewed not as a given event or even a given sequence of events, but as a feedback structure dominated by self-reinforcing positive loops, within which events take place” (p. 338). These types of explanations are based on *mechanisms* as explanatory power and not in simple causal relationships as the source of explanation, even less in (substantialist) causality, i. e. change in singular properties/entities. Furthermore, these hypotheses are developed for each problem consistent with the mental models of the modelers. This is why system dynamics is not committed to specific theories and only to the explanation of problematic behaviors in terms of structure of the model in order to enhance learning and decision-making.

A particular remark must be made. It can be noticed a natural link between mechanism and the idea of “generic structure”. This latter expression has been used in different senses in SD literature. Lane and Smart [58] trace the evolution of this concept – see also [71], they identify three different interpretations. One of these view cannot be connected with mechanism, the one popularized by Senge, e. g. [92,109,110], usually known under the expression “system archetypes”. The lack of computer simulation within this interpretation points at a problem of validity in its scope and claims, as Lane and Smart dis-

cuss [58], e.g. since this perspective skips the possibility of formal computer model building then the relation between structure and behavior is weak and the mentioned approach becomes just a hasty shortcut from problematic behavior to insights and principles, and without the experimental spirit of SD for enhancing learning. But two other notions are relevant. On the one hand, generic structures can be conceived as general models (theories of behavior) of a class of systems that are associated with a domain of application, e.g. urban development, supply chain, economic growth. Lane and Smart label these structures as “canonical situation models”. On the other hand, a generic structure may refer to theories of mathematical structures (feedback loops, levels, rate equations, etc.) that generate corresponding dynamic behaviors, i.e. “systems belong to the same class if they can be represented by the same structure ... This dynamic structure when abstracted from any application domain data defines the class of system” (p. 93 in [58]); therefore they offer transferability of structure across diverse domains. These models can be labeled as “abstracted micro-structures”, e.g. patterns of exponential growth, goal seeking, oscillation, etc. [58,98]. Both interpretations look for establishing a general class of models that formally link structure with behavior and they constitute an important line of research. These developments contribute to different aspects of SD practice; in particular they directly fuel the processes of conceptualization and formal model construction (e.g. see [31,58,98]), and more important, they enhance understanding and the improvement of our mental models as long as we can exploit the powerful idea of having general classes of models, either within a domain of application or across different domains by transferring structures across them.

Consequently, system dynamics explanations can be characterized as *mechanisms*, since there can be found the source of explanatory power. In spite of its causal diagrams, the *explanans*, i.e. that which does the explaining, is based on mechanisms – dynamic hypotheses based on structures; and the problematic behavior is the *explanandum*, i.e. that which is explained. Following Glymour: “Remains, however, a considerable bit of science that sounds very much like explaining, and which perhaps has causal implications, but which does not seem to derive its point, its force, or its interest from the fact that it has something to do with causal relations (or their absence)” (as cited in [p. 212 in 83]). The theories built with SD are essentially structure-based and not content-based (substantialist) explanations i.e. they are not associated with intrinsic properties of objects or entities but with the consequences of processes and activities entrenched in relevant parts of the structure of the model. Recalling Hayek

who identifies explanation with modeling, it is interesting to notice the range of his thought expressed half a century ago and that accurately illustrates SD explanations:

Any model defines a certain range of phenomena which can be produced by the type of situation which it represents. We may not be able directly to confirm that the causal mechanism determining the phenomenon in question is the same as that of the model. But we know that, if the mechanism is the same, the observed structures must be capable of showing some kinds of action and unable to show others (p. 221 in [38]).

A further reminder follows. Since SD was previously identified with instrumentalism, then a mechanism is not to be taken as a description; here a mechanism is an instrument for arranging observations. It should be kept in mind that mechanism is a kind of explanation which refers to epistemological issues.

The popularity of simple causality as the way to characterize the explanation of phenomena contrasts with the assumptions made in SD: structures that generate processes responsible for behavior. This is consistent with the purpose of system dynamics simulation which might be oriented to activities such as theoretical-representations building, articulation and testing in order to learn in and about complex systems [98]. System dynamics uses simulation as a method which is different from the traditional inductive logic of research that deals with single instances which attempts to confirm theories via repeated observation. However, SD does not dismiss presentationalism as it was shown. This should be highlighted as an important and distinctive characteristic of SD. Though there is a commitment with a real world, justification is rooted in social processes and on mental models, and it is also relative to the purpose of the model. Furthermore, the goal is to enhance our decision-making processes by improving our mental models. How can we characterize such method? A short comment follows.

### Simulation and Method

In a plain sense “simulation means driving a model of a system with suitable inputs and observing the corresponding outputs” (p. 23 in [2]). But simulation actually is not just a matter of number crunching. Its scope is broader. And it represents another challenge for philosophy of science. Winsberg [107] illustrates it:

Typically, to a philosopher of science, epistemological issues arise when we try to justify high level the-

oretical claims based on low level data or specific observational reports. But simulation is about starting with theory and working your way down. This kind of epistemology is, to the philosopher of science, a curious beast. It is an epistemology that is concerned with justifying inferences from a theory to its application – an inference that most philosophy of science has assumed is deductive and consequently not in need of justification (p. S447).

What is simulation in SD? It can be affirmed that it is a technique able to represent and test theoretical concepts and not only – in a narrow sense – a tool to just solve mathematical problems. Besides, the emphasis on processes, on patterns of collective action and on the relations between components and its dynamic consequences can be better addressed with simulation because of its capacities to represent these issues with fewer restrictions than other approaches [35]. But there is more. Simulation reflects a very different attitude. This way of inquiry suggests a whole different and new scientific methodology [82,106,107]. Winsberg emphasizes that “simulation represents an entirely new mode of scientific activity – one that lies between theory and experiment ... a form of theory articulation or ‘model building’ (pp. 117, 119 in [106]). Axelrod [2] indeed suggests “a third way” of doing science:

Simulation as a way of doing science can be contrasted with the two standard methods of induction and deduction ... Simulation is a third way of doing science. Like deduction, it starts with a set of explicit assumptions. But unlike deduction, it does not prove theorems. Instead, a simulation generates data that can be analyzed inductively. Unlike typical induction, however, the simulated data comes from a rigorously specified set of rules rather than direct measurement of the real world (p. 24).

Moreover, its strength rests on the capacity for conducting *experiments* [82]. This emphasis on experimentation is the key to understand why this approach is different. Our mental models are nothing more than theoretical models that attribute properties and relations to the systems they represent; the relevance of these theoretical models depends on the purpose of the model. And computer simulation simply permits experiments of these (theoretical) models. This is where the novelty and the power of this methodology are to be found, in the very iterative process of model building and experimentation via simulation. This position contrasts with the traditional prominence of assumed representational capacities of *theories* and *models* where usually the emphasis has been placed, see [107].

But computer simulation has a distinct epistemology [105] that emphasizes *the process of modeling*. The method was demarcated by Forrester [24] in *Industrial Dynamics*:

Simulation consists of tracing through, step by step, the actual flows of orders, goods, and information, and observing the series of new decisions that take place ... This is the counterpart of trying a new policy or organizational structure in the real system... After a simulation run comes interpretation of the results. Did it turn out as expected? If not, why? As the experiment is examined, new questions arise ... This is a process of invention and trial ... Each simulation result teaches, and it also prompts additional questions ... Such experimentation will yield new insights into the characteristics of the system that the model represents (pp. 23, 44–45, 55).

Hence, the method of simulation through continued experimentation is aimed at providing better understanding of the modeled system. As it was mentioned, the method calls attention to the *process* – see also [44,89]. Indeed SD aims at developing a *modeling* culture (consistent with [80]) that gives emphasis to model building as an ongoing dialectic between stakeholders instead of a mapping exercise concerned on the efficacy of the model itself.

Why is fundamental the use of the computer? Perhaps the best answer is provided again by Forrester [26]:

We stress the importance of being explicit about assumptions and interrelating them in a computer model ... The most important difference between the properly conceived computer model and the mental model is in the ability to determine the dynamic consequences when the assumptions within the model interact with one another. The human mind is not adapted to sensing correctly the consequences of a mental model ... The computer model ... is a statement of system structure. It contains the assumptions being made about the system ... Generally, the consequences are unexpected (pp. 213–215).

The shortcomings of our mental models coupled with the complexity of the systems we model lead to the use of the computer. Sterman [101] summarizes these drawback with aspects such our flawed cognitive maps and our erroneous inferences about dynamics. As it was shown, the strength of explanation and understanding in SD is not in the causal models as such; the heart lies in the development of dynamic hypotheses with the use of simulation in order to enhance our understanding and our decision-mak-



ing processes. Explanations are posed under the notion of mechanism; but this is an iterative process that seeks a central aim: to improve our own theories about the world.

### Future Directions

There are several aspects to develop based on this reflection on various assumptions behind system dynamics. The central argument was built around the position known as presentationalism. This stance integrates and informs many of the debates on related subjects, e.g. validation, and it characterizes the initial purposes and assumptions held in the field. However, there are a number of lines to emphasize and to develop.

We have focused on the role of the idea of “mental model” for the practice of SD. Yet, it can be seen utilization of SD models under assumptions which are closer to a naive realism that seem to ignore the purposes of enhancing understanding and learning processes. This article should help to underscore that system dynamics is less naive – and hence more powerful – when we recognize a presentationalist stance which means that our theories about the world are just that, *theories* based on our experience and on the operations of our own mind. We can improve these theories with the use of system dynamics, that is, making our assumptions about systems explicit and using simulation as a method for enhancing understanding and for developing explanations that guide our processes of systems design. This recognition includes the relative nature of justification of knowledge held in SD and the emphasis on mechanism as a powerful way to develop explanations about complex systems.

It has been introduced mechanism as the way to characterize the particular type of explanation pursued in SD. The explanatory force does not rest on causal relations as such but on the structures – physical and decision rules aspects – and on the dynamic processes and activities that explain change. The idea of mechanism is shaped in SD under the expression “generic structure”. The focus on understanding behavior in terms of abstract structures is a central line of inquiry. This article underlines the importance of developing such a line of research since it has been located at the core of the kind of knowledge that SD provides. The issue of unification to provide understanding of diverse phenomena is a definitive step in the way to assert that the field progresses as long as broader range of phenomena may be explained with the same mechanism. Should be the advance of system dynamics assessed by the progress in this type of study? Behind this discussion there are major and provocative issues that arise to be developed.

The long debate of qualitative and quantitative modeling is informed by this characterization. It is clear that the issue of explanation compels theorists and practitioners to ask themselves what is the kind of explanation they are pursuing and if such explanations are enough and satisfying; it is worthy to ask for qualitative modeling what kind of understanding it gives and how it can be characterized, in other words, to give an account of *explanans* and *explanandum*. Another line to develop might be oriented around the following question: what would be *essential* criteria for comparing different arrangements or modes of organization in order to identifying them as belonging to the same type of mechanism? The identification and search of mechanisms becomes a powerful heuristic for guiding the modeling process.

There are promising suggestions for philosophy of science as well. SD offers guidelines, e.g. can the ways in which system dynamicists work provide meaningful insights, or even concrete accounts, for the philosophical unresolved issue of explanation? The mechanism depicted in system dynamics proposes a kind of explanation that goes beyond the received view based on causation. A related question is whether explanation must always follow a deductive path; the classic models of Hempel, e.g. [41], emphasized the condition of deduction and general laws for having an explanation; however, the explanation in SD is not framed under a deductive schema from universal covering laws, instead it can be conceived as a sort of abductive reasoning based on the understanding of the dynamics of the model as a way to understand the actual behavior it accounts for. A further issue is that in spite of the lack of universality, i.e. no universal laws, SD models aim to provide understanding for a diverse range of phenomena that might share relevant influential structures and similar associated behaviors, that is, it accounts for regularities in order to unify them in a certain kind of explanation under the same explanation. This situation insinuates a flavor of paradox because of the traditional rigid association of universal laws with the explanation of regularities, but in SD there are no general laws though the aim is to explain general regularities. The study of *models* and computer simulation – instead of abstract theories and traditional methodologies – is an additional indication that SD suggests for philosophy of science, including the emphasis on the modeling process.

### Bibliography

1. Aronson JL (1988) Testing for Convergent Realism. PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol One: Contributed Papers. University of Chicago Press, Chicago, pp 188–193



2. Axelrod R (1997) Advancing the art of simulation in the social science. In: Conte R, Hegselmann R, Terna P (eds) *Simulating Social Phenomena*. Springer, Berlin, pp 21–40
3. Barlas Y, Carpenter S (1990) Philosophical roots of model validation: two paradigms. *Syst Dyn Rev* 6:148–166
4. Bartley III WW (1987) Philosophy of biology versus philosophy of physics. In: Radnitzky G, Bartley III WW (eds) *Evolutionary epistemology, rationality, and the sociology of knowledge*. Open Court, La Salle, pp 7–45
5. Bechtel W, Richardson RC (1993) *Discovering complexity: Decomposition and localization as strategies in scientific research*. Princeton University Press, Princeton
6. Benjamin AC (1941) Modes of scientific explanation. *Philos Sci* 8:486–492
7. Berger R (1998) Understanding science: Why causes are not enough. *Philos Sci* 65:306–332
8. Berkeley G (1948–1957) *The works of George Berkeley, Bishop of Cloyne*. Thomas Nelson and Sons, London
9. Black M (1934) The principle of verifiability. *Analysis* 2:1–6
10. Blackmore J (1979) On the inverted use of the terms ‘Realism’ and ‘Idealism’ among scientists and historians of science. *Br J Philos Sci* 30:125–134
11. Bowman AA (1916) Kant’s phenomenalism in its relation to subsequent metaphysics. *Mind, New Ser* 25:461–489
12. Brandon RN (1984) Grene on mechanism and reductionism: more than just a side issue. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol II: Symposia and Invited Papers*. University of Chicago Press, Chicago, pp 345–353
13. Brown HI (1990) Charniak on scientific realism. *Br J Philos Sci* 41:415–427
14. Burrell G, Morgan G (1979) *Sociological paradigms and organizational analysis*. Heinemann, London
15. Chapin JP (1941) Idealism and its relation to science. *Philosophy of Science* 8:142–146
16. Checkland P, Pidd M, Morecroft J (2004) Working ideas, insights for systems modeling: The Broader community of systems thinkers. In: Kennedy M, Winch G, Langer R, Rowe J, Yanni J (eds) *Proceedings of the 22nd International Conference of the System Dynamics Society*, Keble College, University of Oxford, England. System Dynamics Society, Albany
17. Coyle G (2000) Qualitative and quantitative modelling in system dynamics: some research questions. *Syst Dyn Rev* 16:225–244
18. Coyle RG (1979) *Management system dynamics*. Wiley, Chichester
19. Deetz S (1996) Describing differences in approaches to organization science: Rethinking Burrell and Morgan and their legacy. *Organ Sci* 7:191–207
20. Dickson M (1995) An empirical reply to empiricism: Protective measurements opens the door for quantum realism. *Philos Sci* 62:122–140
21. Doyle JK, Ford DN (1998) Mental models concepts for system dynamics research. *Syst Dyn Rev* 14:3–29
22. Doyle JK, Ford DN (1999) Mental models concepts revisited: some clarifications and a reply to Lane. *Syst Dyn Rev* 15:411–415
23. Flood R, Jackson M (1991) *Creative problem solving*. Wiley, Chichester
24. Forrester JW (1961) *Industrial Dynamics*. Press MIT, Cambridge
25. Forrester JW (1971) *Principles of Systems*. Wright-Allen Press, Cambridge
26. Forrester JW (1975) Counterintuitive behavior of social systems. In: *Collected papers of Jay W. Forrester*. Wright-Allen Press, Cambridge, pp 211–244
27. Forrester JW (1975) *Industrial Dynamics: A Major breakthrough for decision makers*. In: *Collected papers of Jay W Forrester*. Wright-Allen Press, Cambridge, pp 1–29
28. Forrester JW (1975) *Industrial Dynamics – After the first decade*. In: *Collected papers of Jay W. Forrester*. Wright-Allen Press, Cambridge, pp 133–150
29. Forrester JW (1985) “The” model versus a modeling “process”. *Syst Dyn Rev* 1:133–134
30. Forrester JW (1987) Lessons from system dynamics modeling. *Syst Dyn Rev* 3:136–149
31. Forrester JW (1994) System dynamics, systems thinking, and soft OR. *Syst Dyn Rev* 10:245–256
32. Forrester JW (2003) Dynamic models of economic systems and industrial organizations. *Syst Dyn Rev* 19:331–345
33. Fuller S (1994) Retrieving the point of the realism-instrumentalism debate: Mach vs. Planck on science education policy. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association, vol One: Contributed Papers*. University of Chicago Press, Chicago, pp 200–208
34. Gershenson C, Aerts D, Edmonds B (2007) *Worldviews, science and us*. Philosophy and complexity. World Scientific, Singapore
35. Gilbert N, Troitzsch KG (1999) *Simulation for the social scientist*. Open University Press, Buckingham
36. Glennan SS (2002) Rethinking mechanistic explanation. *Philos Sci* 69:342–353
37. Guyer P (1983) Kant’s intentions in the refutation of idealism. *Philos Rev* 92:329–383
38. Hayek FA (1955) Degrees of explanation. *Br J Philos Sci* 6:209–225
39. Hayek FA (1942) Scientism and the study of society, Part I. *Economica, New Ser* 9:267–291
40. Hayek FA (1943) Scientism and the study of society, Part II. *Economica, New Ser* 10:34–63
41. Hempel CG, Oppenheim P (1948) Studies in the logic of explanation. *Philos Sci* 15:135–175
42. Hesselow G (1981) Causality and determinism. *Philos Sci* 48:591–605
43. Hitchcock CR, Salmon WC (2000) Statistical explanation. In: Newton-Smith WH (ed) *A Companion to the philosophy of science*. Blackwell Publishers, Malden, pp 470–479
44. Homer JB (1996) Why we iterate: scientific modeling in theory and practice. *Syst Dyn Rev* 12:1–19
45. Hume D (1740) *A treatise of human nature*. Oxford University Press, Oxford
46. Humphreys P (2000) Causation. In: Newton-Smith WH (ed) *A companion to the philosophy of science*. Blackwell Publishers, Malden, pp 31–40
47. Hunter B (1992) Empiricism. In: Dancy J, Sosa E (eds) *A Companion to Epistemology*. Blackwell Publishers, Oxford, pp 110–115
48. Jackson M (1991) *Systems methodology for the management sciences*. Plenum Press, New York
49. Jackson M (2003) *Systems thinking: Creative holism for managers*. Wiley, Chichester

50. Jobe EK (1985) Explanation, causality, and counterfactuals. *Philos Sci* 52:357–389
51. Kleindorfer GB, Ganeshan R (1993) The philosophy of science and validation in simulation. In: Evans GW, Mollaghasemi M, Russell EC, Biles WE (eds) *Proceedings of the 1993 Winter Simulation Conference*. IEEE, Piscataway, pp 50–57
52. Kukla A (1995) Scientific realism and theoretical unification. *Analysis* 55:230–238
53. Lane D (1994) With a little help from our friends: How system dynamics and soft OR can learn from each other. *Syst Dyn Rev* 10:101–134
54. Lane D (1999) Friendly amendment: A commentary on Doyle and Ford's proposed re-definition of 'mental model'. *Syst Dyn Rev* 15:185–194
55. Lane D (2000) Should system dynamics be described as a 'Hard' or 'Deterministic' systems approach? *Syst Res Behav Sci* 17:3–22
56. Lane D (2001) *Rerum cognoscere causas: Part I – How do the ideas of system dynamics relate to traditional social theories and the voluntarism/determinism debate?* *Syst Dyn Rev* 17:97–118
57. Lane D (2001) *Rerum cognoscere causas: Part II – Opportunities generated by the agency/structure debate and suggestions for clarifying the social theoretic position of system dynamics.* *Syst Dyn Rev* 17:293–309
58. Lane D, Smart C (1996) Reinterpreting 'generic structure': evolution, application and limitations of a concept. *Syst Dyn Rev* 12:87–120
59. Leplin J (1992) Realism and Methodological Change. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, vol Two: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 435–445
60. Leplin J (2000) Realism and Instrumentalism. In: Newton-Smith WH (ed) *A Companion to the Philosophy of Science*. Blackwell Publishers, Malden, pp 393–401
61. Luna-Reyes LF, Andersen DL (2003) Collecting and analyzing qualitative data for system dynamics: methods and models. *Syst Dyn Rev* 19:271–296
62. Machamer P, Darden L, Craver CF (2000) Thinking about mechanisms. *Philos Sci* 67:1–25
63. Meadows DH (1980) The Unavoidable A Priori. In: Randers J (ed) *Elements of the system dynamics method*. Productivity Press, Cambridge, pp 23–57
64. Milford K (1994) In pursuit of rationality. A note on Hayek's *The Counter-Revolution of Science*. In: Birner J, van Zijp R (eds) *Hayek, Co-ordination and Evolution*. Routledge, London, pp 323–340
65. Mojtahedzadeh M, Andersen D, Richardson GP (2004) Using digest to implement the pathway participation method for detecting influential system structure. *Syst Dyn Rev* 20:1–20
66. Morrison M (1988) Reduction and realism. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, vol One: Contributed Papers. University of Chicago Press, Chicago, pp 286–293
67. Morrison M (1990) Unification, realism and inference. *Br J Philos Sci* 41:305–332
68. Newton-Smith WH (1988) Modest realism. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, vol Two: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 179–189
69. Newton-Smith WH (2000) Explanation. In: Newton-Smith WH (ed) *A companion to the philosophy of science*. Blackwell Publishers, Malden/Oxford, pp 127–133
70. Newton-Smith WH (2000) Hume. In: Newton-Smith WH (ed) *A companion to the philosophy of science*. Blackwell Publishers, Malden, pp 165–168
71. Paich M (1985) Generic structures. *Syst Dyn Rev* 1:126–132
72. Pettit P (1992) Realism. In: Dancy J, Sosa E (eds) *A companion to epistemology*. Blackwell Publishers, Oxford, pp 420–424
73. Popper K (1963) *Conjectures and refutations. The growth of scientific knowledge*. Routledge and Kegan Paul, London
74. Popper K (1968) *The Logic of Scientific Discovery*. Hutchinson, London
75. Psillos S (1996) Scientific realism and the "pessimistic induction". *Proceedings of the Biennial Meeting of the Philosophy of Science Association*, Part I: Contributed Papers. University of Chicago Press, Chicago, pp S306–S314
76. Radnitzky G (1987) In defense of self-applicable critical rationalism. In: Radnitzky G, Bartley III WW (eds) *Evolutionary epistemology, rationality, and the sociology of knowledge*. Open Court, La Salle, pp 279–312
77. Rescher N (1992) Idealism. In: Dancy J, Sosa E (eds) *A companion to epistemology*. Blackwell Publishers, Oxford, pp 187–191
78. Richardson GP (1991) *Feedback thought in social science and systems theory*. Pegasus Communications, Waltham
79. Richardson GP (1996) Problems for the future of system dynamics. *Syst Dyn Rev* 12:141–157
80. Richardson KA (2002) On the limits of bottom-up computer simulation: Towards a nonlinear modeling culture. In: Sprague RH Jr (ed) *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*. IEEE
81. Roberts EB (1978) System dynamics – An introduction. In: Roberts EB (ed) *Managerial applications of system dynamics*. Pegasus Communications Inc., Waltham
82. Rohrlich F (1990) Computer simulation in the physical sciences. *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, vol Two: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 507–518
83. Ruben D (1990) *Explaining explanation*. Routledge, London
84. Runciman WG (1969) What is structuralism? *Br J Sociol* 20:253–265
85. Russell B (1912) *The problems of philosophy*. Oxford University Press, Oxford
86. Russell B (1948) *Human knowledge: Its scope and limits*. Simon and Schuster, New York
87. Salmon W (1992) Explanation. In: Dancy J, Sosa E (eds) *A companion to epistemology*. Blackwell Publishers, Oxford, pp 129–132
88. Salmon WC (2000) Logical Empiricism. In: Newton-Smith WH (ed) *A companion to the philosophy of science*. Blackwell Publishers, Malden, pp 233–242
89. Schaffernicht M (2006) Detecting and monitoring change in models. *Syst Dyn Rev* 22:73–88
90. Schlagel RH (1981) Contextualistic realism. *Philos Phenomenol Res* 41:437–451
91. Sellars W (1976) Is scientific realism tenable? *PSA: Proceedings of the Biennial Meeting of the Philosophy of Science Association*, vol Two: Symposia and Invited Papers. University of Chicago Press, Chicago, pp 307–334

92. Senge P (1990) *The fifth discipline: The art and practice of the learning organization*. Doubleday, New York
93. Sidgwick H, Caird E (1880) Kant's refutation of idealism. *Mind, New Ser* 5:111–115
94. Sinha D (1963) Phenomenology and positivism. *Philos Phenomenol Res* 23:562–577
95. Smith DW (1982) The realism in perception. *Noûs* 16:42–55
96. Spender J-C (1996) Making knowledge the basis of a dynamic theory of the firm. *Strateg Manag J* 17:45–62
97. Stace WT (1944) Positivism. *Mind, New Ser* 53:215–237
98. Sterman J (2000) *Business dynamics. Systems thinking and modeling for a complex world*. McGraw-Hill, Boston
99. Tabery JG (2004) Synthesizing activities and interactions in the concept of a mechanism. *Philos Sci* 71:1–15
100. Taube M (1937) Positivism, science, and history. *J Philos* 34:205–210
101. Turbayne CM (1955) Kant's refutation of dogmatic idealism. *Philos Q* 5:225–244
102. Vanderminden P (2006) System dynamics – A field of study, a methodology or both. In: Größler A, Rouwette E, Langer R, Rowe J, Yanni J (eds) *24th International Conference of The System Dynamics Society*. Radboud University Nijmegen, System Dynamics Society, Albany
103. Vásquez M, Liz M, Aracil J (1996) Knowledge and reality: some conceptual issues in system dynamics modeling. *Syst Dyn Rev* 12:21–37
104. Weimer W (1999) Hayek's approach to the problems of complex phenomena: an introduction to the theoretical psychology of The Sensory Order. In: Boettke P (ed) *The Legacy of Friedrich von Hayek*, vol II. Edward Elgar Publishing Limited, Cheltenham, pp 200–244
105. Winsberg E (1999) Sanctioning models: The epistemology of simulation. *Sci Context* 12:275–293
106. Winsberg E (2003) Simulated experiments: Methodology for a virtual world. *Philos Sci* 70:105–125
107. Winsberg E (2001) Simulations, models, and theories: Complex physical systems and their representations. *Philos Sci* 68:S442–S454
108. Wolstenholme EF (1990) *System enquiry*. Wiley, Chichester
109. Wolstenholme EF (2003) Towards the definition and use of a core set of archetypal structures in system dynamics. *Syst Dyn Rev* 19:7–26
110. Wolstenholme EF (2004) Using generic system archetypes to support thinking and modelling. *Syst Dyn Rev* 20:341–356

## System Regulation and Design, Geometric and Algebraic Methods in

ALBERTO ISIDORI<sup>1</sup>, LORENZO MARCONI<sup>2</sup>

<sup>1</sup> Department of Informatica e Sistemistica, University of Rome, La Sapienza, Italy

<sup>2</sup> Center for Research on Complex Automated Systems (CASY), University of Bologna, Bologna, Italy

### Article Outline

[Glossary](#)

[Definition](#)

[Introduction](#)

[The Generalized Tracking Problem](#)

[The Steady-State Behavior of a System](#)

[Necessary Conditions for Output Regulation](#)

[Sufficient Conditions for Output Regulation](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Exosystem** A dynamical system modeling the set of all exogenous inputs (command/disturbances) affecting a controlled plant.

**Internal model** A model of the exogenous inputs (command/disturbances) affecting a controlled plant, embedded in the interior of the controller.

**Generalized tracking problem** The problem of designing a controller able to asymptotically track/reject any exogenous command/disturbance in a fixed set of functions.

**Observer** A device designed to asymptotically track the state of a dynamical system on the basis of measured observations.

**Steady state** A family of behaviors, in a dynamical system, that are asymptotically approached, as actual time tends to infinity or as initial time tends to minus infinity.

### Definition

A central problem in control theory is the design of feedback controllers so as to have certain outputs of a given plant *to track* prescribed reference trajectories. In any realistic scenario, this control goal has to be achieved in spite of a good number of phenomena which would cause the system to behave differently than expected. These phenomena could be endogenous, for instance parameter variations, or exogenous, such as additional undesired inputs affecting the behavior of the plant. In numerous design problems, exogenous inputs are not available for measurement, nor are known ahead of time, but rather can only be seen as unspecified members of a given family of functions. Embedding a suitable “internal model” of such a family in the controller is a design strategy that has proven to be quite successful in handling uncertainties in the controlled plant as well as in the exogenous inputs.

### Introduction

The problem of controlling the output of a system so as to achieve asymptotic tracking of prescribed trajectories and/or asymptotic rejection of disturbances is a central

problem in control theory. There are essentially three different possibilities to approach the problem: tracking by dynamic inversion, adaptive tracking, tracking via internal models. Tracking by *dynamic inversion* consists in computing a precise initial state and a precise control input (or equivalently a reference trajectory of the state), such that, if the system is accordingly initialized and driven, its output exactly reproduces the reference signal. The computation of such control input, however, requires “perfect knowledge” of the entire trajectory which is to be tracked as well as “perfect knowledge” of the model of the plant to be controlled. Thus, this type of approach is not suitable in the presence of large uncertainties on plant parameters as well as on the reference signal. *Adaptive* tracking consists in tuning the parameters of a control input computed via dynamic inversion in such a way as to guarantee asymptotic convergence to zero of a tracking error. This method can successfully handle parameter uncertainties, but still presupposes the knowledge of the entire trajectory which is to be tracked (to be used in the design of the adaptation algorithm) and therefore an approach of this kind is not suited to the problem of tracking unknown trajectories. Of course, one might consider the problem of tracking a slowly varying reference trajectory as a stabilization problem in the presence of a slowly varying unknown parameter, but this would, in most cases, yield a very conservative solution.

In most cases of practical interest, the trajectory to be tracked (or the disturbance to be rejected) is not available for measurement. Rather, it is only known that this trajectory is simply an (undefined) member in a set of functions, for instance the set of all possible solutions of an ordinary differential equation. These cases include the classical problem of the set point control, the problem of active suppression of harmonic disturbances of unknown amplitude, phase and even frequency, the synchronization of nonlinear oscillations, and similar others. It is in these cases that *tracking via internal models* proves particularly efficient, in its ability to handle simultaneously uncertainties in plant parameters as well as in the trajectory which is to be tracked.

For linear multivariable systems, tracking problems of this kind (those in which the exogenous commands and/or disturbances are only known to be members in the set of solutions of a given ordinary differential equations) have been addressed in very elegant geometric terms by Davison, Francis, Wonham [8,10,11] and others. In particular, one of the most relevant contributions of [11] was a clear delineation of what is known as *internal model principle*, i. e. the fact that the property of perfect tracking is insensitive to plant parameter variations “only if the controller

utilizes feedback of the regulated variable, and incorporates in the feedback path a suitably reduplicated model of the dynamic structure of the exogenous signals which the regulator is required to process”. Conversely, in a stable-closed loop system, if the controller utilizes feedback of the regulated variable and incorporates an internal model of the exogenous signals, the output regulation property is insensitive to plant parameter variations.

A nonlinear enhancement of this theory, which uses a combination of geometry and nonlinear dynamical systems theory, was initiated in [15,16,19] in the context of solving the problem near an equilibrium, in the presence of exogenous signals which were produced by a Poisson stable system. In particular, in [19] it was shown how the use center manifold theory near an equilibrium determines the necessity of the existence of an internal model whenever one can solve an output regulation problem in spite of (small) parameter uncertainties (see also [3]). Since these pioneering contributions, the theory has experienced a tremendous growth, culminating in the development of design methods able to handle the case of parametric uncertainties affecting the autonomous (linear) system which generates the exogenous signals (such as in [9,23]), the case of nonlinear exogenous systems (such as in [5]), or a combination thereof (as in [22]). The purpose of these notes is to present a self-contained exposition of the fundamentals of these design methods.

## The Generalized Tracking Problem

In this article, we address tracking problems that can be cast in the following terms. Consider a finite-dimensional, time-invariant, nonlinear system described by equations of the form

$$\begin{aligned}\dot{x} &= f(w, x, u) \\ e &= h(w, x) \\ y &= k(w, x),\end{aligned}\tag{1}$$

in which  $x \in \mathbb{R}^n$  is a vector of state variables,  $u \in \mathbb{R}^m$  is a vector of inputs used for *control* purposes,  $w \in \mathbb{R}^s$  is a vector of inputs which cannot be controlled and include *exogenous* commands, exogenous disturbances and model uncertainties,  $e \in \mathbb{R}^p$  is a vector of *regulated* outputs which include tracking errors and any other variable that needs to be steered to 0,  $y \in \mathbb{R}^q$  is a vector of outputs that are available for *measurement* and hence used to feed the device that supplies the control action. The problem is to design a controller, which receives  $y(t)$  as input and produces  $u(t)$  as output, able to guarantee that, in the re-



sulting closed-loop system,  $x(t)$  remains bounded and

$$\lim_{t \rightarrow \infty} e(t) = 0, \quad (2)$$

regardless of what the exogenous input  $w(t)$  actually is.

The ability to successfully address this problem very much depends on how much the controller is allowed to know about the exogenous disturbance  $w(t)$ . In the ideal situation in which  $w(t)$  is available to the controller in real-time, the design problem indeed looks much simpler. This is, though, only an extremely optimistic situation which does not represent, in any circumstance, a realistic scenario. The other extreme situation is the one in which nothing is known about  $w(t)$ . In this, pessimistic, scenario the best result one could hope for is the fulfillment of some prescribed ultimate bound for  $|e(t)|$ , but certainly not a sharp goal such as (2). A more comfortable, intermediate, situation is the one in which  $w(t)$  is only known to belong to a fixed family of functions of time, for instance the family of all solutions obtained from a fixed ordinary differential equation of the form

$$\dot{w} = s(w) \quad (3)$$

as the corresponding initial condition  $w(0)$  is allowed to vary on a prescribed set. This situation is in fact sufficiently distant from the ideal but unrealistic case of perfect knowledge of  $w(t)$  and from the realistic but conservative case of totally unknown  $w(t)$ . But, above all, this way of thinking at the exogenous inputs covers a number of cases of major practical relevance. There is, in fact, abundance of design problems in which parameter uncertainties, reference commands and/or exogenous disturbances can be modeled as functions of time that satisfy an ordinary differential equation.

The control law is to be provided by a system modeled by equations of the form

$$\begin{aligned} \dot{\xi} &= \varphi(\xi, y) \\ u &= \gamma(\xi, y) \end{aligned} \quad (4)$$

with state  $\xi \in \mathbb{R}^v$ . The initial conditions  $x(0)$  of the plant (1),  $w(0)$  of the exosystem (3) and  $\xi(0)$  of the controller (4) are allowed to range over fixed compact sets  $X \subset \mathbb{R}^n$ ,  $W \subset \mathbb{R}^s$  and, respectively,  $\mathcal{E} \subset \mathbb{R}^v$ . All maps characterizing the model of the controlled plant, of the exosystem and of the controller are assumed to be sufficiently differentiable.

The problem which will be studied, known as the *generalized tracking problem* (or *problem of output regulation* or also *generalized servomechanism problem*) is to design a feedback controller of the form (4) so as to obtain

a closed loop system in which all trajectories are bounded and the regulated output  $e(t)$  asymptotically decays to 0 as  $t \rightarrow \infty$ . More precisely, it is required that the composition of (1), (3) and (4), that is the *autonomous* system

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{x} &= f(w, x, \gamma(\xi, k(w, x))) \\ \dot{\xi} &= \varphi(\xi, k(w, x)) \end{aligned} \quad (5)$$

with output

$$e = h(w, x),$$

be such that:

- The positive orbit of  $W \times X \times \mathcal{E}$  is bounded, i. e. there exists a bounded subset  $S$  of  $\mathbb{R}^s \times \mathbb{R}^n \times \mathbb{R}^v$  such that, for any  $(w_0, x_0, \xi_0) \in W \times X \times \mathcal{E}$ , the integral curve  $(w(t), x(t), \xi(t))$  of (5) passing through  $(w_0, x_0, \xi_0)$  at time  $t = 0$  remains in  $S$  for all  $t \geq 0$ .
- $\lim_{t \rightarrow \infty} e(t) = 0$ , uniformly in the initial condition, i. e., for every  $\varepsilon > 0$  there exists a time  $\bar{t}$ , depending only on  $\varepsilon$  and not on  $(w_0, x_0, \xi_0)$ , such that the integral curve  $(w(t), x(t), \xi(t))$  of (5) passing through  $(w_0, x_0, \xi_0)$  at time  $t = 0$  satisfies  $\|e(t)\| \leq \varepsilon$  for all  $t \geq \bar{t}$ .

## The Steady-State Behavior of a System

### Limit Sets

The generalized tracking problem can be seen as the problem of forcing in the plant, by means of an appropriate control input  $u(t)$ , a response  $x(t)$  that asymptotically compensates the effect, on the regulated variable  $e(t)$ , of the exogenous input  $w(t)$ . The classical way in which the problem is addressed for linear, time-invariant systems, when the exosystem is a neutrally stable linear system, is to seek a controller forcing in the associated closed-loop system (5) a (stable) “steady state” behavior entirely contained in the kernel of the map defining the tracking error  $e$ . Thus, it is natural to expect that a similar tool should also be effective in the more general setting considered here. It appears, though, that a rigorous investigation of the concept of “steady state”, beyond the classical domain of linear system theory, had never been fully pursued.

Motivated by the current practice in linear system theory, the “steady state” behavior of a dynamical system can be viewed as a kind *limit* behavior, approached either as the *actual* time  $t$  tends to  $+\infty$  or, alternatively, as the *initial* time  $t_0$  tends to  $-\infty$ . Relevant, in this regard, are certain concepts introduced by G.D. Birkhoff in his classical



1927 essay, where he asserts that “with an arbitrary dynamical system . . . there is associated always a closed set of ‘central motions’ which do possess this property of regional recurrence, towards which all other motions of the system in general tend asymptotically” (see p. 190 in [2]). In particular, a fundamental role is played by the concept of  $\omega$ -limit set of a given point, which is defined as follows. Consider an *autonomous* ordinary differential equation

$$\dot{x} = f(x) \quad (6)$$

with  $x \in \mathbb{R}^n$ ,  $t \in \mathbb{R}$ . It is well known that, if  $f: \mathbb{R}^n \rightarrow \mathbb{R}^n$  is locally Lipschitz, for any  $x_0 \in \mathbb{R}^n$ , the solution of (6) with initial condition  $x(0) = x_0$ , denoted by  $x(t, x_0)$ , exists on some open interval of the point  $t = 0$  and is unique.

Assume, in particular, that  $x(t, x_0)$  is defined for all  $t \geq 0$ . A point  $x$  is said to be an  $\omega$ -limit point of the motion  $x(t, x_0)$  if there exists a sequence of times  $\{t_k\}$ , with  $\lim_{k \rightarrow \infty} t_k = \infty$ , such that

$$\lim_{k \rightarrow \infty} x(t_k, x_0) = x.$$

The  $\omega$ -limit set of a point  $x_0$ , denoted  $\omega(x_0)$ , is the union of all  $\omega$ -limit points of the motion  $x(t, x_0)$ .

It is obvious from this definition that an  $\omega$ -limit point is not necessarily a limit of  $x(t, x_0)$  as  $t \rightarrow \infty$ , as the solution in question may not admit any limit as  $t \rightarrow \infty$ . It happens though, that if the motion  $x(t, x_0)$  is *bounded*, then  $x(t, x_0)$  asymptotically approaches the set  $\omega(x_0)$ . This property is precisely described in what follows [2].

**Lemma 1** Suppose there is a number  $M$  such that  $\|x(t, x_0)\| \leq M$  for all  $t \geq 0$ . Then,  $\omega(x_0)$  is a nonempty compact connected set, invariant under (6). Moreover, the distance of  $x(t, x_0)$  from  $\omega(x_0)$  tends to 0 as  $t \rightarrow \infty$ .

One of the remarkable features of  $\omega(x_0)$ , as indicated in this Lemma, is the fact that this set is *invariant* for (6). Invariance means that for any initial condition  $\bar{x}_0 \in \omega(x_0)$  the solution  $x(t, \bar{x}_0)$  of (6) exists for all  $t \in (-\infty, +\infty)$  and that  $x(t, \bar{x}_0) \in \omega(x_0)$  for all such  $t$ . Put in different terms, the set  $\omega(x_0)$  is filled by motions of (6) which are bounded backward and forward in time. The other remarkable feature is that  $x(t, x_0)$  approaches  $\omega(x_0)$  as  $t \rightarrow \infty$ , in the sense that the distance of the point  $x(t, x_0)$  (the value at time  $t$  of the solution of (6) starting in  $x_0$  at time  $t = 0$ ) from the set  $\omega(x_0)$  tends to 0 as  $t \rightarrow \infty$ .

Since any motion  $x(t, x_0)$  which is bounded in positive time asymptotically approaches the  $\omega$ -limit set  $\omega(x_0)$  as  $t \rightarrow \infty$ , one may be tempted to look, for a system (6) in which *all* motions are bounded in positive time, at the union of the limit sets of all points  $x_0$ , i. e. at the set

$$\Omega = \bigcup_{x_0 \in \mathbb{R}^n} \omega(x_0)$$

and to say that the system is in steady state if its state  $x(t)$  evolves in the (invariant) set  $\Omega$ . There is a major drawback, though, in taking this as definition of “steady state” behavior of a nonlinear system: the convergence of  $x(t, x_0)$  to  $\Omega$  is not guaranteed to be *uniform* in  $x_0$ , even if the latter ranges over a compact set (see, e. g. [7]).

One of the main motivations for looking into the concept of steady state is the aim to *shape* the steady state response of a system to a given (or to a given family of) forcing inputs. But this motivation looses much of its meaning if the time needed to get within an  $\varepsilon$ -distance from the steady state may grow unbounded as the initial state changes (even when the latter is picked within a fixed *bounded* set). In other words, *uniform* convergence to the steady state (which is automatically guaranteed in the case of linear systems) is an indispensable feature to be required in a nonlinear version of this notion. The set  $\Omega$ , the union of all  $\omega$ -limit points of all points in the state space does not have this property of uniform convergence, but there is a larger set which does have this property. This larger set, known as the  $\omega$  limit set of a set, is precisely defined as follows.

Consider again system (6), let  $B$  be a subset of  $\mathbb{R}^n$  and suppose  $x(t, x_0)$  is defined for all  $t \geq 0$  and all  $x_0 \in B$ . The  $\omega$ -limit set of  $B$ , denoted  $\omega(B)$ , is the set of all points  $x$  for which there exists a sequence of pairs  $\{x_k, t_k\}$ , with  $x_k \in B$  and  $\lim_{k \rightarrow \infty} t_k = \infty$  such that

$$\lim_{k \rightarrow \infty} x(t_k, x_k) = x.$$

It is clear from the definition that if  $B$  consists of only one single point  $x_0$ , all  $x_k$ ’s in the definition above are necessarily equal to  $x_0$  and the definition in question reduces to the definition of  $\omega$ -limit set of a point, given earlier. It is also clear from this definition that, if for some  $x_0 \in B$  the set  $\omega(x_0)$  is nonempty, all points of  $\omega(x_0)$  are points of  $\omega(B)$ . In fact, all such points have the property indicated in the definition, if all the  $x_k$ ’s are taken equal to  $x_0$ . Thus, in particular, if all motions with  $x_0 \in B$  are bounded in positive time,

$$\bigcup_{x_0 \in B} \omega(x_0) \subset \omega(B).$$

However, the converse inclusion is not true in general.

The relevant properties of the  $\omega$ -limit set of a set, which extend those presented earlier in Lemma 1, can be summarized as follows [14].

**Lemma 2** Let  $B$  be a nonempty bounded subset of  $\mathbb{R}^n$  and suppose there is a number  $M$  such that  $\|x(t, x_0)\| \leq M$  for all  $t \geq 0$  and all  $x_0 \in B$ . Then  $\omega(B)$  is a nonempty compact

set, invariant under (6). Moreover, the distance of  $x(t, x_0)$  from  $\omega(B)$  tends to 0 as  $t \rightarrow \infty$ , uniformly in  $x_0 \in B$ . If  $B$  is connected, so is  $\omega(B)$ .

Thus, as it is the case for the  $\omega$ -limit set of a point, we see that the  $\omega$ -limit set of a bounded set, being compact and invariant, is filled with motions which exist for all  $t \in (-\infty, +\infty)$  and are bounded backward and forward in time. But, above all, we see that the set in question is uniformly approached by motions with initial state  $x_0 \in B$ , a property that the set  $\Omega$  does not have. Note also that the set of all such trajectories is a “behavior”, in the sense of J.C. Willems [25].

The set  $\omega(B)$ , as shown in the previous Lemma, asymptotically attracts, as  $t \rightarrow \infty$ , all motions that start in  $B$ . Since the convergence to  $\omega(B)$  is uniform in  $x_0$ , it is also true that, whenever  $\omega(B)$  is contained in the interior of  $B$ , the set  $\omega(B)$  is *asymptotically stable*, in the sense of Lyapunov (see [14]).

### The Steady State Behavior of a Nonlinear System

Consider now again system (6), with initial conditions in a closed subset  $X \subset \mathbb{R}^n$ . Suppose the set  $X$  is *positively invariant*, which means that for any initial condition  $x_0 \in X$ , the solution  $x(t, x_0)$  exists for all  $t \geq 0$  and  $x(t, x_0) \in X$  for all  $t \geq 0$ . The motions of this system are said to be *ultimately bounded* if there is a bounded subset  $B$  with the property that, for every compact subset  $X_0$  of  $X$ , there is a time  $T > 0$  such that  $\|x(t, x_0)\| \in B$  for all  $t \geq T$  and all  $x_0 \in X_0$ . In other words, if the motions of the system are ultimately bounded, every motion eventually enters and remains in the bounded set  $B$ .

Suppose the motions of (6) are ultimately bounded and let  $B' \neq B$  be any other bounded subset with the property that, for every compact subset  $X_0$  of  $X$ , there is a time  $T > 0$  such that  $\|x(t, x_0)\| \in B'$  for all  $t \geq T$  and all  $x_0 \in X_0$ . Then, it is easy to check that  $\omega(B') = \omega(B)$ . Thus, in view of the properties described in Lemma 2 above, the following definition can be adopted (see [7]).

**Definition** Suppose the motions of system (6), with initial conditions in a closed and positively invariant set  $X$ , are ultimately bounded. A *steady state* motion is any motion with initial condition  $x(0) \in \omega(B)$ . The set  $\omega(B)$  is the *steady state locus* of (6) and the *restriction* of (6) to  $\omega(B)$  is the *steady state behavior* of (6).

The notion introduced in this way recaptures the classical notion of steady state for linear systems and provides a new powerful tool to deal with similar issues in the case of nonlinear systems.

*Example* In order to see how this notion includes the classical viewpoint, consider an  $n$ -dimensional, single-input, asymptotically stable linear system

$$\dot{z} = Fz + Gu \quad (7)$$

forced by the harmonic input  $u(t) = u_0 \sin(\omega t + \phi_0)$ . A simple method to determine the periodic motion of (7) consists in viewing the forcing input  $u(t)$  as provided by an autonomous “signal generator” of the form

$$\dot{w} = Sw \quad u = Qw$$

in which

$$S = \begin{pmatrix} 0 & \omega \\ -\omega & 0 \end{pmatrix} \quad Q = \begin{pmatrix} 1 & 0 \end{pmatrix}$$

and in analyzing the state state behavior of the associated “augmented” system

$$\begin{aligned} \dot{w} &= Sw \\ \dot{z} &= Fz + GQw. \end{aligned} \quad (8)$$

As a matter of fact, let  $\Pi$  be the unique solution of the Sylvester equation  $\Pi S = F\Pi + GQ$  and observe that the graph of the linear map

$$\begin{aligned} \pi: \mathbb{R}^2 &\rightarrow \mathbb{R}^n \\ w &\mapsto \Pi w \end{aligned}$$

is an invariant subspace for the system (8). Since all trajectories of (8) approach this subspace as  $t \rightarrow \infty$ , the limit behavior of (8) is determined by the restriction of its motion to this invariant subspace.

Revisiting this analysis from the viewpoint of the more general notion of steady state introduced above, let  $W \subset \mathbb{R}^2$  be a set of the form

$$W = \{w \in \mathbb{R}^2: \|w\| \leq c\} \quad (9)$$

in which  $c$  is a fixed number, and suppose the set of initial conditions for (8) is  $W \times \mathbb{R}^n$ . This is in fact the case when the problem of evaluating the periodic response of (7) to harmonic inputs whose amplitude does not exceed a fixed number  $c$  is addressed. The set  $W$  is compact and invariant for the upper subsystem of (8) and, as it is easy to check, the  $\omega$ -limit set of  $W$  under the motion of the upper subsystem of (8) is the subset  $W$  itself.

The set  $W \times \mathbb{R}^n$  is closed and positively invariant for the full system (8) and, moreover, since the lower subsystem of (8) is a linear asymptotically stable system driven by a bounded input, it is immediate to check that the motions

of system (8), with initial conditions taken in  $W \times \mathbb{R}^n$ , are ultimately bounded. As a matter of fact, any bounded set  $B$  of the form

$$B = \{(w, z) \in \mathbb{R}^2 \times \mathbb{R}^n : w \in W, \|z - \Pi w\| \leq d\}$$

in which  $d$  is any positive number, has the property indicated in the definition of ultimate boundedness. It is easy to check that

$$\omega(B) = \{(w, z) \in \mathbb{R}^2 \times \mathbb{R}^n : w \in W, z = \Pi w\},$$

i. e.  $\omega(B)$  is the graph of the restriction of the map  $\pi$  to the set  $W$ . The restriction of (8) to the invariant set  $\omega(B)$  characterizes the steady state behavior of (7) under the family of all harmonic inputs of fixed angular frequency  $\omega$ , and amplitude not exceeding  $c$ .

*Example* A similar result, namely the fact that the *steady state locus* is the *graph* of a map, can be reached if the “signal generator” is any nonlinear system, with initial conditions chosen in a compact invariant set  $W$ . More precisely, consider an augmented system of the form

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{z} &= Fz + Gq(w), \end{aligned} \quad (10)$$

in which  $w \in W \subset \mathbb{R}^r$ ,  $x \in \mathbb{R}^n$ , and assume that: (i) all eigenvalues of  $F$  have negative real part, (ii) the set  $W$  is a compact set, invariant for the the upper subsystem of (10).

As in the previous example, the  $\omega$ -limit set of  $W$  under the motion of the upper subsystem of (10) is the subset  $W$  itself. Moreover, since the lower subsystem of (10) is a linear asymptotically stable system driven by the bounded input  $u(t) = q(w(t, w_0))$ , the motions of system (10), with initial conditions taken in  $W \times \mathbb{R}^n$ , are ultimately bounded.

It is easy to check that the steady state locus of (10) is the graph of the map

$$\begin{aligned} \pi : W &\rightarrow \mathbb{R}^n \\ w &\mapsto \pi(w), \end{aligned}$$

defined by

$$\pi(w) = \lim_{T \rightarrow \infty} \int_{-T}^0 e^{-F\tau} Gq(w(\tau, w)) d\tau. \quad (11)$$

To see why this is the case, pick any initial condition  $(w_0, z_0)$  for (10) on the graph of  $\pi$  and compute the solution  $z(t)$  of the lower equation of (10) by means of the classical variation of constants formula, to obtain

$$z(t) = e^{Ft} z_0 + \int_0^t e^{F(t-\tau)} Gq(w(\tau, w_0)) d\tau$$

Since by hypothesis  $z_0 = \pi(w_0)$ , using (10) one obtains

$$\begin{aligned} z(t) &= e^{Ft} \int_{-\infty}^0 e^{-F\tau} Gq(w(\tau, w_0)) d\tau \\ &\quad + \int_0^t e^{F(t-\tau)} Gq(w(\tau, w_0)) d\tau \\ &= \int_{-\infty}^t e^{F(t-\tau)} Gq(w(\tau, w_0)) d\tau \\ &= \int_{-\infty}^0 e^{-F\theta} Gq(w(\theta + t, w_0)) d\theta \\ &= \int_{-\infty}^0 e^{-F\theta} Gq(w(\theta, w(t, w_0))) d\theta \\ &= \pi(w(t, w_0)) = \pi(w(t)) \end{aligned}$$

which proves the invariance of the graph of  $\pi$  for (10). It is deduced from this that any point of the graph of  $\pi$  is necessarily a point of the steady state locus of (10). To complete the proof of the claim it remains to show that no other point of  $W \times \mathbb{R}^n$  can be a point of the steady state locus. But this is a straightforward consequence of the fact that  $F$  has eigenvalues with negative real part.

There are various ways in which the result discussed in the previous example can be generalized. For instance, it can be extended to describe the steady state response of a nonlinear system

$$\dot{z} = f(z, u) \quad (12)$$

in the neighborhood of a locally exponentially stable equilibrium point. To this end, suppose that  $f(0, 0) = 0$  and that the matrix

$$F = \left[ \frac{\partial f}{\partial z} \right] (0, 0)$$

has all eigenvalues with negative real part. Then, it is well known (see e.g. p. 275 [13]) that it is always possible to find a compact subset  $Z \subset \mathbb{R}^n$ , which contains  $z = 0$  in its interior and a number  $\sigma > 0$  such that, if  $\|z_0\| \in Z$  and  $\|u(t)\| \leq \sigma$  for all  $t \geq 0$ , the solution of (12) with initial condition  $z(0) = z_0$  satisfies  $\|z(t)\| \in Z$  for all  $t \geq 0$ . Suppose that the input  $u$  to (12) is produced, as before, by a signal generator of the form

$$\begin{aligned} \dot{w} &= s(w) \\ u &= q(w) \end{aligned} \quad (13)$$

with initial conditions chosen in a compact invariant set  $W$  and, moreover, suppose that,  $\|q(w)\| \leq \sigma$  for all  $w \in W$ . If this is the case, the set  $W \times Z$  is positively invariant for

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{z} &= f(z, q(w)), \end{aligned} \quad (14)$$

and the motions of the latter are ultimately bounded, with  $B = W \times Z$ . The set  $\omega(B)$  may have a complicated structure but it is possible to show, by means arguments similar to those which are used in the proof of the Center Manifold theorem, that if  $Z$  and  $B$  are small enough the set in question can still be expressed as the graph of a map  $z = \pi(w)$ . In particular, the graph in question is precisely the center manifold of (14) at  $(0, 0)$  if  $s(0) = 0$  and the matrix

$$S = \left[ \frac{\partial s}{\partial w} \right] (0)$$

has all eigenvalues on the imaginary axis.

A common feature of the examples discussed above is the fact that the steady state locus of a system of the form (14) can be expressed as the graph of a map  $z = \pi(w)$ . This means that, so long as this is the case, a system of this form has a *unique* well defined *steady state response* to the input  $u(t) = q(w(t))$ . As a matter of fact, the response in question is precisely  $z(t) = \pi(w(t))$ . Of course, this may not always be the case and *multiple* steady state responses to a given input may occur. In general, the following property holds.

**Lemma 3** *Let  $W$  be a compact set, invariant under the flow of (13). Let  $Z$  be a closed set and suppose that the motions of (14) with initial conditions in  $W \times Z$  are ultimately bounded. Then, the steady state locus of (14) is the graph of a set-valued map defined on the whole of  $W$ .*

### Necessary Conditions for Output Regulation

Taking advantage of the notions introduced in the previous section, we are now in a position to highlight some general properties that any controller that solves a problem of output regulation must necessarily have. Recall that, as defined earlier, the problem of output regulation is solved if, in the composite system (5):

- the positive orbit of  $W \times X \times \mathcal{E}$  is bounded,
- $\lim_{t \rightarrow \infty} e(t) = 0$ , uniformly in the initial condition.

The notions introduced in the previous section are instrumental to prove the following, elementary – but fundamental – result, which is a nonlinear enhancement of a Lemma of [10] on which all the theory of output regulation for linear systems is based.

**Lemma 4** *Suppose the positive orbit of  $W \times X \times \mathcal{E}$  is bounded. Then*

$$\lim_{t \rightarrow \infty} e(t) = 0$$

*if and only if*

$$\omega(W \times X \times \mathcal{E}) \subset \{(w, x, \xi): h(w, x) = 0\}. \quad (15)$$

It is seen from this simple result that the problem of output regulation can be simply cast as the problem of *shaping the steady state locus of the closed loop system*, in such a way that property (15) holds.

To proceed with the analysis in a more concrete fashion, we consider from now on the special case in which the controlled plant (4) is modeled by equations *in normal form*

$$\begin{aligned} \dot{z} &= f_0(w, z) + f_1(w, z, e_1)e_1 \\ \dot{e}_1 &= e_2 \\ &\vdots \\ \dot{e}_{r-1} &= e_r \\ \dot{e}_r &= q(w, z, e_1, \dots, e_r) + b(w, z, e_1, \dots, e_r)u \\ e &= e_1 \\ y &= \text{col}(e_1, \dots, e_r), \end{aligned} \quad (16)$$

with state  $(z, e_1, \dots, e_r) \in \mathbb{R}^{n-r} \times \mathbb{R}^r$ , control input  $u \in \mathbb{R}$ , regulated output  $e \in \mathbb{R}$ , measured output  $y \in \mathbb{R}^r$ . The functions  $f_0(\cdot)$ ,  $f_1(\cdot)$ ,  $q(\cdot)$ ,  $b(\cdot)$ ,  $s(\cdot)$  in (16) and (3) are assumed to be at least continuously differentiable. It is also assumed that

$$b(w, z, e_1, \dots, e_r) \neq 0 \quad \forall (w, z, e_1, \dots, e_r).$$

The initial conditions of (16) range on a set  $Z \times E$ , in which  $Z$  is a fixed *compact* subset of  $\mathbb{R}^{n-r}$  and  $E = \{(e_1, \dots, e_r) \in \mathbb{R}^r: |e_i| \leq c\}$ , with  $c$  a fixed number.

Suppose that a controller of the form (4) solves the problem of output regulation. Then Lemma 4 applies and, since  $e = e_1$ , we deduce that the steady state locus of the closed loop system (5) is necessarily a subset of the set of all states in which  $e_1 = 0$ . This being the case, it is seen from the form of the equations (16) that, when the closed loop system (5) is in steady state, necessarily also

$$e_2 = e_3 = \dots = e_r = 0.$$

As a consequence, the following conclusions hold:

- The steady state locus  $\omega(W \times Z \times E \times \mathcal{E})$  of the closed-loop system is a subset of the set  $\mathbb{R}^s \times \mathbb{R}^{n-r} \times \{0\} \times \mathbb{R}^v$ .
- The restriction of the closed-loop system to its steady state locus  $\omega(W \times Z \times E \times \mathcal{E})$  reduces to

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{z} &= f_0(w, z) \\ \dot{\xi} &= \varphi(\xi, 0). \end{aligned} \quad (17)$$

- For each  $(w, z, 0, \dots, 0, \xi) \in \omega(W \times Z \times E \times \mathcal{E})$

$$0 = q(w, z, 0, \dots, 0) + b(w, z, 0, \dots, 0)\gamma(\xi, 0). \quad (18)$$

The prior analysis implicitly assumes that the positive orbit of  $W$  under the flow of exosystem is bounded, i. e. that the motions of the exosystem asymptotically approach the its own steady state locus  $\omega(W)$ . In principle,  $\omega(W)$  may differ from  $W$  but there is no loss of generality in assuming from the very beginning that the two sets coincide. After all, the problem in question is a problem concerning how the closed-loop system behaves in steady state and there is no special interest in considering exosystems that are not “in steady state”. We make this assumption precise as follows.

**Assumption (i)** *The compact set  $W$  is invariant for (3).*

With this in mind we observe that, by Lemma 3, if the positive orbit of  $W \times Z \times E \times \mathcal{E}$  under the flow of (5) is bounded, then  $\omega(W \times Z \times E \times \mathcal{E})$  is the graph of a (possibly set-valued) map defined on the whole of  $W$ . Consider now the set

$$\mathcal{A}_{ss} = \{(w, z) : (w, z, 0, \dots, 0, \xi) \in \omega(W \times Z \times E \times \mathcal{E}), \text{ for some } \xi \in \mathbb{R}\}$$

and define the map

$$u_{ss} : \mathcal{A}_{ss} \rightarrow \mathbb{R} \\ (w, z) \mapsto -\frac{q(w, z, 0, \dots, 0)}{b(w, z, 0, \dots, 0)}.$$

By construction, the set  $\mathcal{A}_{ss}$  is the graph of a (possibly set-valued) map defined on the whole of  $W$ , which is invariant for the dynamics of

$$\begin{aligned} \dot{w} &= s(w) \\ \dot{z} &= f_0(w, z), \end{aligned} \quad (19)$$

that are precisely the zero dynamics of the “augmented system” (3)–(16), while the map  $u_{ss}(\cdot)$  is the control that forces the motion of (3)–(16) to evolve on  $\mathcal{A}_{ss}$ .

With this in mind, the conclusions reached above can be rephrased in the following terms. Suppose that a controller of the form (4) solves the problem of output regulation for (16) with exosystem (3). Then, there exists a (possibly set-valued) map defined on the whole of  $W$  whose graph  $\mathcal{A}_{ss}$  is invariant for the autonomous system (19). Moreover, for each  $(w_0, z_0) \in \mathcal{A}_{ss}$  there is a point  $\xi_0 \in \mathbb{R}^v$  such that the integral curve of (19) issued from  $(w_0, z_0)$  and the integral curve of

$$\dot{\xi} = \varphi(\xi, 0)$$

issued from  $\xi_0$  satisfy

$$u_{ss}(w(t), z(t)) = \gamma(\xi(t)), \quad \forall t \in \mathbb{R}.$$

This is a nonlinear version of the celebrated *internal model principle* of [11].

## Sufficient Conditions for Output Regulation

### The Control Structure

On the basis of the ideas presented in the previous section we proceed now with the construction of a controller that solves the problem of output regulation. The “steady state” features of this controller are those identified at the end of the section, namely this controller has to be able to “generate” all controls of the form  $u_{ss}(w(t), z(t))$  for any “steady state” trajectory  $w(t), z(t)$  of (19). The controller should incorporate a device that generates all such trajectories (the *internal model*), thus making sure that the “appropriate” state-state behavior takes place, and a device guaranteeing that convergence to this specific steady state behavior occurs. It is here that additional assumptions are needed.

Note that, since  $W$  is invariant for  $\dot{w} = s(w)$ , the closed cylinder

$$C := W \times \mathbb{R}^{n-r}$$

is locally invariant for (19). Hence, it is natural regard (19) as a system defined on  $C$  and endow the latter with the subset topology.

**Assumption (ii)** *There exists a bounded subset  $B$  of  $C$  which contains the positive orbit of the set  $W \times Z$  under the flow of (19) and the resulting omega-limit set*

$$\mathcal{A} := \omega(W \times Z)$$

satisfies

$$(w, z) \in C, \quad |(w, z)|_{\mathcal{A}} \leq d_0 \quad \Rightarrow \quad z \in Z \quad (20)$$

where  $d_0$  is a positive number.

While in the analysis of the necessity we have only identified the existence of a compact set (actually, the graph of a map defined on  $W$ ) which is invariant for (19), the new assumption (ii) implies, in its first part, the existence of a compact set  $\mathcal{A}$  (still the graph of a map defined on  $W$ ) which is not only invariant but also uniformly attractive of all trajectories of (19) issued from points of  $W \times Z$ . The second part of the assumption, in turn, guarantees that this set is also stable in the sense of Lyapunov. In the next assumption we strengthen this property by also requiring the



set  $\mathcal{A}$  to be locally exponentially stable (this assumption is useful to straighten the subsequent analysis, but is not essential).

**Assumption (iii)** *There exist  $M \geq 1$ ,  $\lambda > 0$  such that*

$$(w_0, z_0) \in C, \quad |(w_0, z_0)|_{\mathcal{A}} \leq d_0 \quad \Rightarrow \\ |(w(t), z(t))|_{\mathcal{A}} \leq M e^{-\lambda t} |(w_0, z_0)|_{\mathcal{A}}, \quad \forall t \geq 0$$

in which  $(w(t), z(t))$  denotes the solution of (19) passing through  $(w_0, z_0)$  at time  $t = 0$ .

To simplify the exposition, we address the special case in which the controlled system (16) has relative degree 1, and in which the coefficient  $b(w, z, e_1)$  is identically equal to 1. In other words, we consider a system modeled by equations of the form

$$\begin{aligned} \dot{z} &= f_0(w, z) + f_1(w, z, e)e \\ \dot{e} &= q(w, z, e) + u \\ y &= e. \end{aligned} \quad (21)$$

There is no loss of generality in considering a system having this simple form (21) because, as shown for instance in [9,22], the case of a more general system of the form (16) can easily be reduced, by appropriate manipulations, to this one.

For convenience, rewrite the augmented system (3)–(21) as

$$\begin{aligned} \dot{z} &= f_0(z) + f_1(z, e)e \\ \dot{e} &= q_0(z) + q_1(z, e)e + u \end{aligned} \quad (22)$$

having set  $z = (w, z)$ . Consistently let  $Z := W \times Z$  denote the compact set where the initial condition  $z(0)$  is supposed to range. In these notations, assumptions (i) – (ii) – (iii) express the property that, in the autonomous system

$$\dot{z} = f_0(z), \quad (23)$$

the set  $\mathcal{A}$  is asymptotically and locally exponentially stable, with a domain of attraction that contains the set  $Z$ .

Suppose now that the control  $u$  is chosen as  $u = -ke$ . The closed-loop system thus obtained can be regarded as a feedback interconnection of

$$\dot{z} = f_0(z) + f_1(z, e)e \quad (24)$$

viewed as a system with input  $e$  and state  $z$ , and

$$\dot{e} = q_0(z) + q_1(z, e)e - ke \quad (25)$$

viewed as a system with input  $z$  and state  $e$ .

By assumption, system (24) possesses, when  $e = 0$ , an invariant set  $\mathcal{A}$  which is asymptotically and locally exponentially stable, with a domain of attraction that contains the set  $Z$  of all admissible initial conditions. Thus, standard arguments (see, e.g. [6].) can be invoked to claim that, if  $k$  is large enough, all trajectories of the interconnection (24)–(25) with initial conditions in  $Z \times E$  remain bounded and the state  $(z, e)$  can be steered to an arbitrary small neighborhood of the set  $\mathcal{A} \times \{0\}$ . This does not solve the problem at issue, though, because the variable  $e(t)$  is not guaranteed to converge to zero (but only to converge to a neighborhood of zero, whose size can be made arbitrarily small by increasing the gain coefficient  $k$ ). The condition for having  $e(t) \rightarrow 0$  as  $t \rightarrow \infty$  is simply that the “coupling” term  $q_0(z)$  vanishes on the set  $\mathcal{A}$ , but there is no reason for this to occur (see again, e.g. [6]). This is why a more elaborate, internal-model-based, controller is needed.

System (16) being affine in the control input  $u$ , it seems natural to look for a controller having a similar structure, namely a controller of the form

$$\begin{aligned} \dot{\xi} &= \varphi(\xi) + Gv \\ u &= \gamma(\xi) + v \end{aligned} \quad (26)$$

with state  $\xi \in \mathbb{R}^v$ , in which  $v$  is a residual control input, to be eventually chosen as a function of the measured output  $y$ . Here  $\varphi(\cdot)$ ,  $G$  and  $\gamma(\cdot)$  are functions to be determined. We will show in what follows that, if the triplet  $\{\varphi(\xi), G, \gamma(\xi)\}$  possesses what we will define as *asymptotic internal model* property, the choice of the residual control  $v$  in (26) as

$$v = -ke$$

solves the problem of output regulation, provided that the gain coefficient  $k$  is sufficiently high.

### The Internal Model

Controlling this system by means of (26) yields a closed-loop system

$$\begin{aligned} \dot{z} &= f_0(z) + f_1(z, e)e \\ \dot{e} &= q_0(z) + q_1(z, e)e + \gamma(\xi) + v \\ \dot{\xi} &= \varphi(\xi) + Gv \end{aligned} \quad (27)$$

which, regarded as a system with input  $v$  and output  $e$ , has relative degree 1 and zero dynamics given by

$$\begin{aligned} \dot{z} &= f_0(z) \\ \dot{\xi} &= \varphi(\xi) - G[\gamma(\xi) + q_0(z)]. \end{aligned} \quad (28)$$

System (27) can be put in normal form by means of the change of variables

$$\chi = \xi - Ge$$

which yields

$$\begin{aligned}\dot{z} &= f_0(z) + f_1(z, e)e \\ \dot{\chi} &= \varphi(\chi + Ge) - G\gamma(\chi + Ge) - Gq_0(z) - Gq_1(z, e)e \\ \dot{e} &= q_0(z) + q_1(z, e)e + \gamma(\chi + Ge) + v.\end{aligned}\quad (29)$$

Setting  $x = (z, \chi)$ , this system can be further rewritten in the form

$$\begin{aligned}\dot{x} &= f(x) + \ell(x, e)e \\ \dot{e} &= q(x) + r(x, e)e + v\end{aligned}\quad (30)$$

in which

$$f(x) = \begin{pmatrix} f_0(z) \\ \varphi(x) - G[\gamma(x) + q_0(z)] \end{pmatrix}$$

$$q(x) = q_0(z) + \gamma(x)$$

and  $\ell(x, e)$ ,  $r(x, e)$  are suitable continuous functions.

Suppose now that the residual control  $v$  is chosen as  $v = -ke$ . This yields a closed-loop system having a structure similar to the one considered in the previous subsection, namely a feedback interconnection of

$$\dot{x} = f(x) + \ell(x, e)e \quad (31)$$

viewed as a system with input  $e$  and state  $x$ , and

$$\dot{e} = q(x) + r(x, e)e - ke \quad (32)$$

viewed as a system with input  $x$  and state  $e$ . As claimed earlier, a high-gain control on  $e$ , namely a control  $v = -ke$  with large  $k$ , would succeed in steering  $e(t)$  to zero if two conditions are fulfilled:

(P1) the dynamics (28) possesses a compact invariant set which is asymptotically (and locally exponentially) stable, with a domain of attraction that contains the set  $Z \times \mathcal{E}$  of all admissible initial conditions, and

(P2) the function  $q_0(z) + \gamma(\xi)$  vanishes on this invariant set.

These are the properties that will be sought in what follows. Note that the fulfillment of these is determined only by properties of the autonomous system (23) and of the function

$$\rho = q_0(z) \quad (33)$$

which, in the composite system (28), can be viewed as the output of (23) driving a system of the form

$$\dot{\xi} = \varphi(\xi) - G[\gamma(\xi) + \rho]. \quad (34)$$

For convenience, we will say that triplet  $\{\varphi(\xi), G, \gamma(\xi)\}$  is an asymptotic internal model of the pair (23) – (33) if properties (P1) and (P2) are satisfied. In this terminology, we can summarize as follows the conclusion obtained so far.

**Proposition 1** *Pick compact sets  $Z$ ,  $E$  and  $\mathcal{E}$  for the initial conditions of the closed-loop system (3), (21), (26). Assume that (i)-(ii)-(iii) hold and that the triplet  $\{\varphi(\xi), G, \gamma(\xi)\}$  is an asymptotic internal model of (23) – (33). Then there exists  $k^* > 0$  such that for all  $k \geq k^*$  the controller (26) with  $v = -ke$  solves the generalized tracking problem.*

The notion of steady state provides a useful interpretation of the properties in question. In fact, recall that, by assumption, all trajectories of system (23) with initial conditions in  $Z$  asymptotically converge to the compact invariant set  $\mathcal{A}$ , and the latter is also locally exponentially stable. If property (P1) holds, all trajectories of the composite system (28) with initial conditions in  $Z \times \mathcal{E}$  asymptotically converge to the limit set  $\omega(Z \times \mathcal{E})$ . Since (28) is a triangular system, it is readily seen that the set  $\omega(Z \times \mathcal{E})$  is the graph of a set-valued map defined on  $\mathcal{A}$ , i. e. that there exists a map

$$\tau : z \in \mathcal{A} \mapsto \tau(z) \subset \mathbb{R}^p,$$

such that

$$\omega(Z \times \mathcal{E}) = \{(z, \xi) : z \in \mathcal{A}, \xi \in \tau(z)\} = \text{gr}(\tau).$$

The set  $\text{gr}(\tau)$  is the steady state locus of (28) and the restriction of the latter to this invariant set characterizes its steady state behavior. Property (P2), on the other hand, expresses the property that at each point of  $(z, \xi) \in \text{gr}(\tau)$

$$q_0(z) = -\gamma(\xi). \quad (35)$$

Thus, looking again at system (28), it is realized that  $\text{gr}(\tau)$  is in fact invariant for

$$\begin{aligned}\dot{z} &= f_0(z) \\ \dot{\xi} &= \varphi(\xi).\end{aligned}\quad (36)$$

Note that, if the map  $\tau(z)$  is single-valued and  $C^1$ , its invariance for (36) is expressed by the property that

$$\frac{\partial \tau(z)}{\partial z} f_0(z) = \varphi(\tau(z)) \quad \forall z \in \mathcal{A}, \quad (37)$$

while the fact that (35) holds at each point of  $(z, \xi) \in \text{gr}(\tau)$  is expressed by the property that

$$q_0(z) = -\gamma(\tau(z)) \quad \forall z \in \mathcal{A}. \quad (38)$$

### The Design of an Internal Model

As we have seen in the earlier sections, the proposed controller, if the asymptotic internal model property holds, is able to force – in the closed loop system – convergence to a steady state in which the regulated variable is identically zero. As a consequence, the controller solves the generalized tracking problem. It remains to be shown, therefore, how the asymptotic internal model property can be obtained. To this end, it is convenient to observe that the properties required in (P1) and (P2) are quite similar to properties that are usually sought in the design of *state observers*. As a matter of fact it is seen from (37) and (38) that, for each  $z_0 \in \mathcal{A}$ , the function of time

$$\hat{\xi}(t) = \tau(z(t), z_0)$$

which is defined (and bounded) for all  $t \in \mathbb{R}$  satisfies

$$\frac{d\hat{\xi}(t)}{dt} = \varphi(\hat{\xi}(t)) \quad (39)$$

and, moreover

$$\gamma(\hat{\xi}(t)) = -q_0(z(t), z_0).$$

In view of the latter, system (34) can be rewritten in the form

$$\dot{\xi} = \varphi(\xi) + G[\gamma(\hat{\xi}) - \gamma(\xi)] \quad (40)$$

and interpreted as a *copy of the dynamics* (39) of  $\hat{\xi}$  corrected by an “innovation term”  $[\gamma(\hat{\xi}) - \gamma(\xi)]$  weighted by an “output injection gain”  $G$ . This is the classical structure on an *observer* and the requirement in (P1) expresses the property that the difference  $\xi(t) - \hat{\xi}(t)$  (the “observation error”, in our interpretation) should asymptotically decay to zero (with ultimate exponential decay).

This interpretation is at the basis of a number of major recent advances in the design of regulators. In fact, in a number of recent papers, this interpretation has been pursued and, taking into consideration various approaches to the design of nonlinear observers, has lead to effective design methods (see [5,9,22]). Two of such design methods are highlighted in the remaining part of this section.

### The High-Gain Observer as an Internal Model (see [5])

The construction summarized in this section relies upon the following additional hypothesis.

**Assumption (iv)** Suppose there exist an integer  $d > 0$  and a locally Lipschitz function  $f: \mathbb{R}^d \rightarrow \mathbb{R}$  such that, for any

$z_0 \in \mathcal{A}$ , the solution  $z(t)$  of passing through  $z_0$  at time  $t = 0$  is such that the function  $\rho(t) := q_0(z(t))$  satisfies

$$\rho^{(d)}(t) = f(\rho(t), \rho^{(1)}(t), \dots, \rho^{(d-1)}(t))$$

for all  $t \in \mathbb{R}$ .

Let  $\tau: W \times \mathbb{R}^{n-1} \rightarrow \mathbb{R}^d$  be the map defined as

$$\tau(z) := \text{col}(q_0(z), L_{f_0} q_0(z), \dots, L_{f_0}^{d-1} q_0(z)) \quad (41)$$

and let  $f_c: \mathbb{R}^d \rightarrow \mathbb{R}$  be a  $C^1$  function with compact support which agrees with  $f(\cdot)$  on  $\tau(\mathcal{A})$ . Then, it easy to check that the properties indicated in (37) and (38) are fulfilled by choosing

$$\varphi(\xi) = \begin{pmatrix} \xi_2 \\ \vdots \\ \xi_d \\ f_c(\xi_1, \xi_2, \dots, \xi_d) \end{pmatrix}, \quad \gamma(\xi) = \xi_1. \quad (42)$$

Comparing this construction with the earlier remarks we observe, in particular, that system

$$\begin{aligned} \dot{z} &= f_0(z) \\ \rho &= q_0(z) \end{aligned} \quad (43)$$

is immersed into a system which is uniformly observable, in the sense of [12] (even though system (43) might not have had such a property). It is precisely this that makes it possible to choose  $G$  in such a way that the property indicated in (P1) can be achieved.

As a matter of fact, the property in question is achieved by choosing

$$G = D_g \begin{pmatrix} c_0 \\ \vdots \\ c_{d-1} \end{pmatrix}$$

where  $D_g = \text{diag}(g, g^2, \dots, g^d)$ ,  $g$  is a design parameter, and the  $c_i$ 's are such that the polynomial  $\lambda^d + c_0 \lambda^{d-1} + \dots + c_{d-1} = 0$  is Hurwitz, as formally proved in Lemmas 1 and 2 of [5] to which the interested reader is referred for details.

It is worth noting that the assumption in question clearly covers the interesting (and widely addressed in the recent past literature, see [15]) case in which the function  $f(\cdot)$  is linear, namely the case in which (43) is immersed into a linear observable system. In this case, although the choice indicated above is clearly still valid, a more direct way of designing the regulator is to use  $f(\cdot)$  instead of  $f_c(\cdot)$  in the definition of  $\varphi(\xi)$ , and simply choose  $G$  in such a way that  $\dot{\xi} = \varphi(\xi) - G\gamma(\xi)$  is a stable linear system.

**The Andrieu-Praly's Observer as an Internal Model** (see [22]) In this subsection we exploit certain results of the theory presented in [1] to weaken (and, to some extent, suppress) the Assumption (iv) presented at the beginning of the earlier subsection.

Let  $(F, G) \in \mathbb{R}^{d \times d} \times \mathbb{R}^{d \times 1}$  be a controllable pair and set

$$\varphi(\xi) = F\xi + G\gamma(\xi), \quad (44)$$

with  $\gamma: \mathbb{R}^d \rightarrow \mathbb{R}$  a continuous function to be determined later. If this is the case, the composite system (28) becomes

$$\begin{aligned} \dot{z} &= f_0(z) \\ \dot{\xi} &= F\xi - Gq_0(z), \end{aligned} \quad (45)$$

which is precisely a system of the form (9), considered in the Example. If the matrix  $F$  is Hurwitz, and we restrict  $z$  to belong to the set  $\mathcal{A}$ , this system has a well defined steady state behavior, which is the graph of the map

$$\tau(z) = \int_{-\infty}^0 e^{-Fs} Gq_0(z(s, z)) ds. \quad (46)$$

As shown in Example, the graph in question is invariant for system (45), is asymptotically (and locally exponentially) stable and with a domain of attraction that coincides with  $W \times \mathbb{R}^d$ . Moreover, it can also be shown that there exists a number  $\ell > 0$  such that, if the eigenvalues of  $F$  have real part which is less  $\ell$ , the map (46) is  $C^1$  (see, e. g. [22]). If this is the case, to say that the graph of (46) is invariant for (45) is equivalent to say that

$$\frac{\partial \tau}{\partial z} f_0(z) = F\tau(z) - Gq_0(z) \quad \forall z \in \mathcal{A}. \quad (47)$$

This being the case, it is immediate to check that properties (P1) and (P2) will be fulfilled if a function  $\gamma(\xi)$  can be found that renders (38) satisfied. As a matter of fact, bearing in mind (44), condition (37) becomes

$$\frac{\partial \tau}{\partial z} f_0(z) = F\tau(z) - G\gamma(\tau(z)) \quad \forall z \in \mathcal{A}$$

which, if condition (38) holds, reduces to (47). This, shows that a triplet having the asymptotic internal model property can be found if a function  $\gamma(\cdot)$  exists which satisfies (38). It is here that the dimension  $d$  of the pair  $(F, G)$  plays a role, as formalized in the next proposition whose proof can be found in [22].

**Proposition 2** *Suppose*

$$d \geq 2(s + n - r) + 2.$$

*Then for almost all choices (see [22] for details) of a controllable pair  $(F, G)$ , with  $F$  a Hurwitz matrix whose eigenvalues have real part which is less than  $\ell$ , the map (46) satisfies*

$$\tau(z_1) = \tau(z_2) \Rightarrow q_0(z_1) = q_0(z_2).$$

*As a consequence there exist a continuous map  $\gamma: \tau(\mathcal{A}) \rightarrow \mathbb{R}$  such that*

$$q_0(z) = -\gamma(\tau(z)) \quad \forall z \in \mathcal{A}. \quad (48)$$

The map  $\tau(\cdot)$  in (46) is defined only on  $\mathcal{A}$ , but is not difficult to extend it to a  $C^1$  map defined on the whole set  $W \times \mathbb{R}^{n-r}$ , as shown in [22]. Also the map  $\gamma(\cdot)$  that makes (48) true can be extended to the whole  $\mathbb{R}^d$ , but this extension is only known to be continuous.

We have shown in this way that the existence of triplet  $\psi(\xi), G, \gamma(\xi)$  which has the internal model property can always be achieved, so long as the integer  $d$  is large enough. This result shows that the general design procedure outlined earlier in the article is always applicable (so long as the standing hypotheses (i)–(ii)–(iii) are applicable). From the constructive viewpoint, though, it must be observed that the result indicated in Proposition PR1 is only an existence result and that the function  $\gamma(\xi)$ , whose existence is guaranteed, is only known to be continuous. Obtaining continuous differentiability of such  $\gamma(\xi)$  and a constructive procedure are likely to require further hypotheses, which should be in any case weaker than Assumption (iv) considered earlier, whose study is subject of current investigation.

## Future Directions

One of the basic hypotheses of the theory described in the previous sections is that the zero dynamics of the augmented system consisting of the controlled plant and of the exosystem possess a compact attractor which is also locally asymptotically stable. This assumption, with an acceptable abuse of terminology, is usually referred to as the “minimum-phase” property. Another standing hypothesis is that the regulated variable coincides with the measured variable. The future directions of the research in this area are aimed at the removal of these assumptions. There are several directions in which the problem can be tackled. One way is the use of control structures by means of which the controlled system (plus a part of the controller) can be interpreted with systems having a different zero dynamics (in particular a zero dynamics possessing a compact attractor). This is the nonlinear equivalent of certain design procedures for linear systems based on the assignment of zeros. This technique has proven to be power-

ful in the stabilization of certain classes of nonlinear systems (see [18]) and is expected to be successful, with appropriate enhancements, in the design of regulators. The analysis of the necessary conditions for output regulation has also shown that, if the generalized tracking problem is solvable, the inverse dynamics of the augmented system consisting of the controlled plant and of the exosystem possesses a compact invariant set to which all initial conditions are asymptotically controllable (by means of the regulated variable viewed as a control). This set is not necessarily asymptotically stable (as it would be under the “minimum-phase” hypothesis) but could be asymptotically stabilizable (by either full-state or measurement-based feedback). Thus, another direction in which the research will evolve, in the development of control schemes for possibly non “minimum-phase” nonlinear systems, is based on the exploitation of the (weaker) assumption that in the zero dynamics of the augmented system there is a compact set that can be made invariant and asymptotically stable by means of feedback. This will yield a design procedure in which a virtual control (either full-state or measurements-based) is designed to stabilize that compact invariant set. The (possibly dynamic) controller obtained in this way will then be embedded into a regulator designed according to the principles developed in this article.

## Bibliography

- Andrieu V, Praly L (2006) On the existence of a Kazantsev-Kravars/Luenberger observer. *SIAM J Contr Optim* 45:432–456
- Birkhoff GD (1927) *Dynamical systems*. American Mathematical Society, Providence
- Byrnes CI, Delli Priscoli F, Isidori A, Kang W (1997) Structurally stable output regulation of nonlinear systems. *Automatica* 33:369–385
- Byrnes CI, Isidori A (2003) Limit Sets, Zero dynamics and internal models in the problem of nonlinear output regulation. *IEEE Trans Autom Contr* 48:1712–1723
- Byrnes CI, Isidori A (2004) Nonlinear internal models for output regulation. *IEEE Trans Autom Contr* 49:2244–2247
- Byrnes CI, Isidori A, Praly L (2003) On the asymptotic properties of a system Arising in non-equilibrium theory of output regulation, preprint of the Mittag-Leffler Institute. 18, Stockholm
- Isidori A, Byrnes CI (2007) The steady-state response of a nonlinear system: ideas, tools and applications, preprint
- Davison EJ (1976) The robust control of a servomechanism problem for linear time-invariant multivariable systems. *IEEE Trans Autom Contr* AC-21:25–34
- Delli Priscoli F, Marconi L, Isidori A (2006) A new approach to adaptive nonlinear regulation. *SIAM J Contr Optim* 45:829–855
- Francis BA (1977) The linear multivariable regulator problem. *SIAM J Contr Optim* 14:486–505
- Francis BA, Wonham WM (1976) The internal model principle of control theory. *Automatica* 12:457–465
- Gauthier JP, Kupka I (2001) *Deterministic observation theory and applications*. Cambridge University Press, Cambridge
- Hahn W (1967) *Stability of motions*. Springer, New York
- Hale JK, Magalhães LT, Oliva WM (2002) *Dynamics in infinite dimensions*. Springer, New York
- Huang J, Lin CF (1994) On a robust nonlinear multivariable servomechanism problem. *IEEE Trans Autom Contr* 39:1510–1513
- Huang J, Rugh WJ (1990) On a nonlinear multivariable servomechanism problem. *Automatica* 26:963–972
- Isidori A (1995) *Nonlinear Control Systems*. Springer, London
- Isidori A (2000), A tool for semiglobal stabilization of uncertain non-minimum-phase nonlinear systems via output feedback. *IEEE Trans Autom Contr* AC-45:1817–1827
- Isidori A, Byrnes CI (1990) Output regulation of nonlinear systems. *IEEE Trans Autom Contr* 25:131–140
- Isidori A, Marconi L, Serrani A (2003) *Robust autonomous guidance: An internal model-based approach*. Springer, London
- Khalil H (1994) Robust servomechanism output feedback controllers for feedback linearizable systems. *Automatica* 30:587–599
- Marconi L, Praly L, Isidori A (2006) Output stabilization via nonlinear luenberger observers. *SIAM J Contr Optim* 45:2277–2298
- Serrani A, Isidori A, Marconi L (2001) Semiglobal nonlinear output regulation with adaptive internal model. *IEEE Trans Autom Contr* 46:1178–1194
- Teel AR, Praly L (1995) Tools for semiglobal stabilization by partial state and output feedback. *SIAM J Control Optim* 33:1443–1485
- Willems JC (1991) Paradigms and puzzles in the theory of dynamical systems, *IEEE Transaction on Automatic Control* 36:259–294

## Systems Biology of Human Immunity and Disease

JARED C. ROACH

Seattle Childrens Hospital, Seattle, USA

### Article Outline

[Glossary](#)

[Definition of the Subject](#)

[Introduction](#)

[Challenges and Solutions for Human Systems Biology](#)

[Human-Specific Problems](#)

[Examples](#)

[Future Directions](#)

[Bibliography](#)

### Glossary

**Complex disease** A disease that has an etiology inconsistent with the simple models of genetic inheritance proposed by Gregor Mendel. A synonym for “complex disease” is “non-Mendelian disease.” Type 2 diabetes is



an example of a complex disease, as are most common diseases. The term “complex phenotype” is employed analogously to describe non-disease properties of individuals, such as height. Typically, complex diseases are influenced by multiple genetic and environmental factors. For most diseases, little is known about either of these sets of factors.

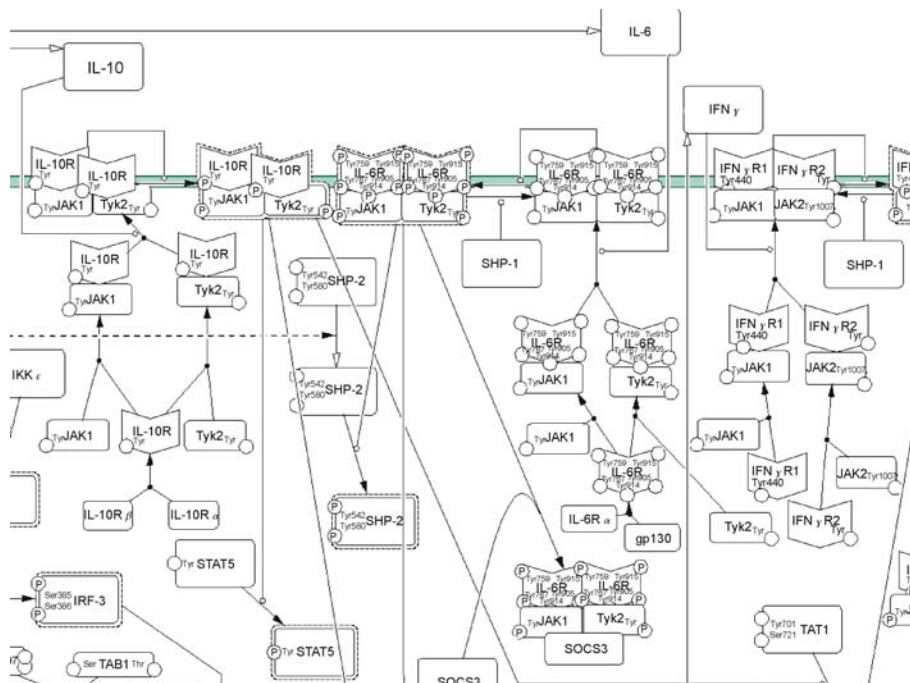
**Nuclear regulation** Many signals are integrated and processed on segments of DNA that are near or adjacent to genes. These segments of DNA are called cis-regulatory elements. Proteins that bind cis-regulatory elements are called transcription factors. The informational output of cis-regulatory signal integration is the rate of production of messenger RNA (mRNA) encoded by that gene. A major goal for many human systems biology projects is to understand key nuclear regulatory networks.

**Regulatory network** A paradigm for modeling information flow within a biological system, such as a cell. Elements of the system that can have multiple states are encoded as nodes; modes of communication between these elements are encoded as edges. A typical node would be a protein that might have multiple states related to its level of phosphorylation or cellular location. A typical edge would represent the catalytic effect of one protein upon another's state of phosphorylation.

**Signaling pathway** Signals typically reach cells through the binding of molecules to receptors on the cell surface. Over the course of seconds to minutes, the receptor interacts with other proteins or small molecules, changing their structure, concentration, or intracellular location. These changes can in turn cause other changes. One common ultimate effect is an alteration in transcription factor binding to target genes, in turn causing a change in mRNA levels. The term “signaling pathway” pre-dates modern systems biology, and represents a paradigm for information flow that is largely unidirectional and linear. Systems biology replaces this paradigm with that of a “regulatory network.”

### Definition of the Subject

Systems biology is the derivation of emergent properties of a multicomponent biological system through the construction of quantitative predictive models. These predictive models are typically formulated as networks. A portion of such a model is shown in Fig. 1. Human systems biology is the practice of systems biology to elucidate emergent properties of humans or any subcomponent of the human body, such as a cell. Human systems present challenges and opportunities not found in other organisms.



Systems Biology of Human Immunity and Disease, Figure 1

Example of a portion of a network model of a macrophage signaling network. This network was drafted with aid of CellDesigner software [34], and is reproduced with permission of The Systems Biology Institute (period)

Most problems in human systems biology are motivated by a desire to understand, predict, prevent, ameliorate, or cure a human disease. The human immune system is a mediator for most human diseases, and is the key system mediating autoimmune and infectious diseases. Many problems in human systems biology focus on the immune system.

As defined, human systems biology is a subset of systems biology. The subject merits special consideration largely because of medical importance. The human subset of systems biology is not an exceptionally distinct subset compared to other reasonable subsets, such as mammalian systems biology. However, it is fairly distinct from systems biology as applied to unicellular and invertebrate organisms.

## Introduction

The difference between human systems biology and the practice of systems biology in other organisms is one of philosophy and emphasis [1]. The need to improve human health drives human systems biology. The emphasis of general systems biology is the development of technological, methodological, and algorithmic approaches to science. Human systems biology exploits these developments, but tends to be more focused on clinical endpoints. It is acceptable to approach these endpoints with incomplete, partial, hybrid, or modified versions of systems biology. Because these approaches include elements of systems biology, such as the analysis of high-throughput data, they are labeled as systems biology, even though they may not include all elements standard to general systems biology, such as the use of multiple orthologous data sets [5].

The needs of modern medicine set the context for human systems biology. From a data-driven point of view, there have been three major epochs of medicine: (1) observational medicine, (2) evidence-based medicine, and (3) predictive and personalized medicine [51]. These categories overlap to some extent, but they correspond fairly well to distinct periods of time. Observational medicine was prevalent from the dawn of civilization until early in the twentieth century. It is characterized by personal experience and oral tradition. An example and highlight of observational medicine was the discovery of digitalis as a treatment for dropsy, now known as edema. The discovery is credited in 1776 to William Withering, an English doctor who lived outside of London. Dr. Withering noticed that a patient with dropsy had been cured by a folk practitioner using a preparation from the foxglove plant. It is likely that this discovery had been made many times prior to the mid-eighteenth century, but lack of or-

ganized methods for communication and data handling kept the knowledge fragmented and incomplete. Dr. Withering was rich and connected. He was a member of the Lunar Society, a group of scientists founded by Matthew Boulton and Erasmus Darwin, the grandfather of Charles. These connections, foreshadowing those available to modern scientists through the internet, enabled the knowledge of foxglove to be preserved, distributed, and improved.

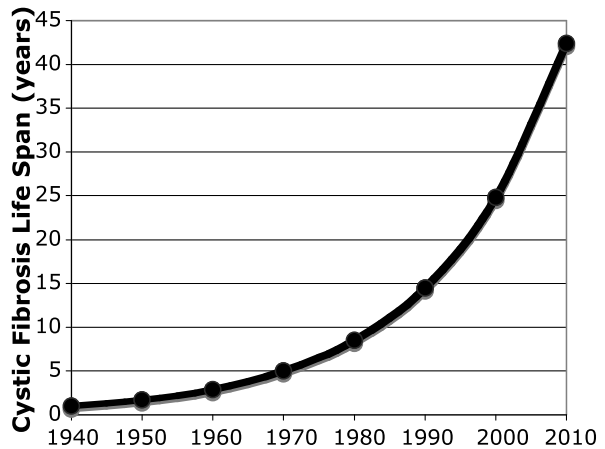
By the early twentieth century, incremental improvements to medical knowledge became increasingly common, and evidence-based medicine flourished. Advancements were enabled primarily by: (1) advances in communication, such as inexpensive publishing by printing presses, (2) chemistry, permitting the purification of active ingredients from botanical medicines, and (3) the nascent field of statistics. Statistics enabled legitimacy and falsifiability of medical facts. The first major application of statistics to medicine was the work of John Snow in 1843 to correlate water sources with cholera outbreaks. This correlation, and many others since, enabled preventive medicine. Chemistry, and later other fields of science and engineering, enabled reductionism. For example, reductionism enabled the identification of digitalis as the component of foxglove responsible for its effects on the cardiovascular system. The culmination of evidence-based medicine was the Framingham Heart Study circa 1948. Thousands of people from the town of Framingham and their descendants were observed longitudinally. Many health parameters were recorded. The Framingham study provided much of our modern knowledge concerning heart disease risk, including effects of diet, exercise, and aspirin. The Framingham Heart Study gave birth to the term “risk factor” and bolstered the recognition of prevention as the most cost-effective form of medicine. Prior to preventive medicine as a philosophy, a heart attack would be viewed as the beginning of a relationship between patient and doctor. With preventive medicine, the relationship would begin at birth or even before, and focus on modifying risk factors to prevent an attack from ever occurring. Data processing and analysis for such studies became increasingly sophisticated and required computers. Progress in evidence-based medicine was linked to the simultaneous revolution in computational power. In particular, the Framingham Heart Study benefited from scientific and sociological momentum generated in World War II. Operations Research was largely born in the Battle of the Atlantic. The US population was willing to participate in large studies with society as the primary beneficiary. Command-and-control bureaucracies existed to coordinate large projects. Projects requiring the collaboration of hundreds of medi-

cal professionals were favorably viewed by researchers and their funding agencies.

By the end of the Human Genome Project at the beginning of the twenty-first century, limitations of the previous century's approach to evidence-based medicine were beginning to be recognized and solutions to these limitations were envisioned. Statistics, exemplified by linear regression models and chi-squared tests, required increasingly large cohorts to identify weaker risk factors. Knowledge gained from large cohorts was applicable only to populations (such as "all white males") and not individuals (such as the person reading this article). Studies became very expensive and could take decades or even centuries to elucidate correlations between causes and effects. The stage was set for the era of predictive and personalized medicine. A major approach to developing knowledge for this era of data-driven medicine has been and will be human systems biology.

In 1892, William Osler observed, "If it were not for the great variability among individuals medicine might be a science, not an art." Historically, the art of medicine was to personalize by trial and error. In the absence of personalization, a patient was prescribed a treatment based on population-based statistics. This is still the norm. If the treatment fails, the next best treatment from a population-inferred list will be prescribed. This process will repeat until the patient or doctor tires of it or dies, money becomes limiting, or effective treatment is achieved. Personalized medicine aims to match the best treatment for an individual without such perilous trial-and-error—and to turn Dr. Osler's art into a science. Previously, data for personalized models was based on orally conveyed family history, together gender and a crude estimate of genetic heritage, or race. Such personalization was tailored to subpopulations of people. 21st-century personalized medicine will develop predictions from an individual's genes and personal developmental and exposure history. Interventions will be tailored to that individual, not to a subpopulation to which the individual belongs. Systems-biology research creates the data and knowledge required by this approach. In particular, models are created based on the understanding of molecular mechanisms. Emergent properties arise from holistic analysis of the interactions in the modeled system. Each individual can be modeled based on their genetic risks, environmental exposures and their responses, biomarkers, and health assessments. Recommendations for particular medications, doses, interventions, and lifestyle changes will be personalized [12].

Expectations for modern research may be portrayed very highly in the popular press, with hopes for miracle cures. These may arise, but the most profound changes in



**Systems Biology of Human Immunity and Disease, Figure 2**

Exponential rise in expected life span for cystic fibrosis. A fairly low exponent, 5.5% in this case, can lead to dramatic advances in human health over time. The goals of human systems biology are to drive such exponential advances

medicine will most likely come from steady incremental advances. Research in all epochs of medicine has produced incremental advances. Cumulatively, these advances have enabled vast gains in human health (Fig. 2). Future incremental advances are expected throughout the epoch of personalized and preventive medicine begins.

The phrase "systems biology" was first used early in the twentieth century to describe analyses of ecosystems, but the phrase never became popular until the turn of 21st century with the establishment of several institutes and research groups focused on systems-biology research and using the term "systems biology" to describe that research. The timing of this surge in interest was almost entirely due to the completion of the Human Genome Project and the need to develop research methodologies appropriate for the post-genomic era. The complete genome sequence provided a near-comprehensive "parts list" of genes. Coinciding approximately with the Human Genome Project, high-throughput technologies for measuring other important parts of human systems were developed, including technologies for measuring proteins and transcripts [18]. Neologisms with the suffix "omics" describing these new fields of analysis have become popular, including "proteomics" and "transcriptomics." Other factors driving modern systems biology have been (1) the tremendous boost to human collaboration and information sharing provided by the internet and (2) the conceptualization of biology as an information science. The digital nature of the genome has led to a belief that all of the individual components and interactions of biological

systems are ultimately knowable, can be represented in data structures, and are amenable to computational analysis. The elucidation of the digital genome is a triumph of reductionism that has energized the pursuit of emergent properties. In the near future, complete sequences of the genomes of many humans will be known, further facilitating correlations between genotype and phenotype.

Modern systems biology focuses on systems at the cellular or subcellular level because systems of these complexities are considered to be neither too trivial to study nor too complex to comprehend. Some systems biology projects involve systems that are either more or less complex. In particular, human systems biology tends to focus on larger, more complex systems. For most problems of medical interest, it is generally anticipated that clinically useful predictions will require modeling at the multicellular level or at the level of the entire organism. The need to focus on very complex systems and to make clinically useful predictions underlies fundamental differences between the conduct of systems biology in humans and in other organisms.

There are tens of thousands of genes, small molecules, and proteins in a typical cellular system. Properties of these components and of the entire system need to be studied under many different conditions and dynamically over time periods ranging from seconds to years. There are several important generalizations that can be made about the techniques required to study such systems. Systems biology requires: (1) the analysis of high-throughput data sets, (2) the analysis of multiple qualitatively distinct different types of data (“orthologous data”), (3) a multidisciplinary collaborative team capable of both data generation (“bench work”) and data analysis (“computational biology”). Also, because systems biology makes predictions, its practice is necessarily iterative: a model is built, predictions are tested, discrepancies between predictions and measurements are used to refine the model, predictions of the new model are tested, and so on until the incremental utility of the model refinements becomes small [47].

The application of the above approach to science is not new; it is merely new to biology. Other disciplines have used collaborative research employing multiple high-throughput data sources to produce models and emergent predictions. For example, in meteorology, a daily weather report is a simple prediction derived from such a process. One of the utilities of the phrase “systems biology” is that it brings attention to the methodological differences between this approach and the more prevalent reductionist approaches in biology during the twentieth century. The coining of the phrase thus helps catalyze a sociological change in biology.

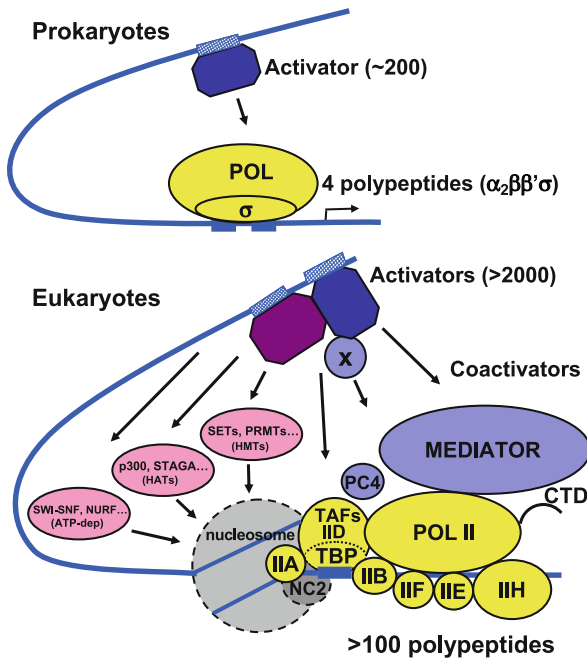
A number of disciplines that might semantically be considered to fit the definition of systems biology are not usually implied by users of the phrase because these disciplines were well established prior to the completion of the Human Genome Project. These include ecology, physiology, and biochemistry. In these fields, research may involve formulation of hypotheses as networks, and prediction of emergent properties of multicomponent systems [26].

### Challenges and Solutions for Human Systems Biology

Human systems biology requires asking questions that are distinct from typical questions asked of prokaryotic systems. These distinct questions require distinct methodologies to answer [30].

There are many challenges for human systems biology. Underlying network models are far more complex than those of prokaryotes [43]. There are more components, or network nodes, of nearly every type. For example, there are more genes, as discussed in the accompanying chapter on genome complexity. Each gene has more cis-regulatory elements. On average, any given gene has more functions and products. These products are primarily mRNAs. On average, any given mRNA has more functions and products. These products are primarily proteins. Eukaryotic cells have more cellular compartments, including mitochondria, lysosomes, phagosomes, a nucleus, endoplasmic reticulum, and secretory granules. Humans have many more cells, and more cell types, and they are organized into different tissues [6]. Furthermore, there is no guarantee that insights gained from one human tissue will be easily extrapolatable to other human tissues; the same protein may play a very different role. In addition to many more nodes in human networks, there are many more interactions. In particular, the number and complexity of interactions operating on cis-regulatory elements of eukaryotic genes is greater (Fig. 3) [42,48].

Not only are human systems much more complex than other systems, but it is much harder to acquire data from these systems. Many experiments possible in other organisms are impossible in humans due to ethical concerns. Such concerns prevent the intentional creation of humans with identical genomes. Humans have very long life spans that essentially rule out acquiring data over the life span of particular individuals, particularly for systems biology, which requires iteration. Projects that might take weeks in a prokaryotic context would take many centuries if the experimental strategies were directly applied to human problems. The cost of performing experiments on human



**Systems Biology of Human Immunity and Disease, Figure 3**  
Eukaryotic versus prokaryotic cis-regulation. The number of factors involved in regulating expression of a eukaryotic gene is at least an order of magnitude greater than for a prokaryotic gene. The complexity of the interactions is also greater. Reprinted from [42]

systems also tends to be much higher than in other systems.

As a consequence of the difficulty of research in mammals, whether that research is reductionist or holistic, on average for any given node or edge of a human network there is less prior information available than there would be for a simple model organism. In many cases, a literature or database search for information pertaining to a human gene will yield either no information or information restricted to a few cell types under only a few conditions [11]. For example, the target genes and sequences of most human transcription factors are completely unknown, leading to substantial uncertainties in models of human nuclear regulatory networks.

There are many solutions to these challenges. Some solutions remain visionary or require technologies that are in development, such as the ability to sequence an individual's genome for less than \$1,000. If the sole goal of human research was knowledge for the sake of knowledge, then the most economically efficient approach to most human systems biology problems would be to wait for these technologies to be fully developed before applying them directly to disease research. However, medicine carries a per-

sonal urgency not found in most other areas of research. Delays in research can be measured in human life and well being. So in addition to investing in technology and methodology, human systems biology must work with the tools at hand.

Because so much less is known about the basic elements of human systems, a continued focus on the individual nodes and edges is important. Human systems biology projects do not displace traditional reductionist biology projects, but rather rely on them. A key fundamental question to be asked of all possible nodes in a network is that of membership: "Does this node belong in the network?" The same question can be asked of each possible edge. The various qualities of the network components also need to be determined. These include directionality of directed edges and rate constants for parts of the network modeled dynamically. If the most important nodes and edges in a network can be determined, future reductionist research can be appropriately directed to expand information on these network components. The complexity of human systems, the difficulty of automatically parsing the literature, and the high false positive rate in many human high-throughput data sets demands increased expert curation at all steps in the process. The sheer volume of the data makes such a call for human curation seemingly futile, but currently without such curation, most models will not make useful medical predictions. Appropriately directed projects will employ automated tools to reduce the burden of curation. For this purpose some of the most useful tools will be interfaces that present curators with information in forms that are easy to comprehend and integrate manually and that also permit decisions and rationales to be quickly recorded. The visual computer interfaces imagined for the 2002 science fiction movie "Minority Report" serve to stimulate ideas of what such interfaces might become.

Human systems biologists must also model simpler, or smaller, networks than would be appropriate for a simple model organism. Although somewhat paradoxical, this need arises from the intense need for data curation and the difficulty of acquiring high-quality data. The more complex the systems, the greater the number of free parameters. For larger networks, the lack of experimental data prevents these parameters from being constrained, and so no statistical confidence can be placed in predictions. Therefore, a reasonable approach for building a predictive network in a human system is to start very small. In some cases, networks with as few as two nodes and a single connecting edge can form a reasonable nucleus for the iterative process of building a systems biology network. Although such a network would be too trivial to discuss in



a prokaryotic setting, it might be capable of making a clinically useful prediction and so would merit consideration in a human disease setting.

One possible approach to reducing the network complexity is to focus on only a few key nodes. Another is to model systems at grosser levels [9]. Thus rather than representing each protein as a node, one might group a set of proteins as a conceptual “module,” and only represent the module as a node in the network. However, this approach requires exceptional care and curation, as it becomes hard to define and even harder to measure quantitative properties of modules [30]. One approach to developing hypotheses related to the functionality of modules is to perform comparative network analysis between different vertebrates [8]. Such analysis requires at least a moderate understanding of a network in at least two systems, which is currently rare, but may become more available over the next few decades.

It is easy for researchers to be distracted by the vast quantity of high-throughput data available, and to yield to reporting analyzes that are the result of automated in silico predictions of algorithms that often cannot be tested against gold standards. Human curation is often viewed as subjective, as it is hard to provide numerical statistics associated with such curation. Numerical statistics, on the other hand, are fairly easy to provide with output from automated algorithms, and these outputs are considered objective. From a medical point of view, however, the criterion of utility must trump evaluation of research based on simple ratio of subjectivity to objectivity. Medical objectivity must be evaluated at a deeper level of objectivity: quantitative measures of improved health care. Frequently, the output of automated algorithms is indeed objective, but is also nonsense. Examples of such nonsense are often manifested as very large networks consisting of hundreds or thousands of nodes and edges. Such networks are said to resemble hairballs and have been termed “ridiculomes.” These networks have little if any predictive capability [32]. A weather report, as published in a daily newspaper, offers an example of a prediction in a useful form. Systems biology predictions should aim to be more like weather reports and less like hairballs.

Despite the large number of obstacles facing human research, there are a few paths to knowledge that work very well in humans. These derive largely from the ability of human research subjects to intellectually contribute to a research project. This is especially true for phenotypes of higher cognitive function. The most successful examples of human-specific research projects are family and genome-wide association studies. Volunteers with specific phenotypes bring themselves to the attention of

researchers, permitting the genetic difference between affected and unaffected individuals to be determined. This is the most powerful and productive method that human systems biologists currently have for identifying the genes that are key nodes in human disease systems.

Given the fairly large number of differences in the state of knowledge and in the constraints on how research might be performed, the genre of questions asked by human systems biologists differs markedly from questions asked in other systems. Whereas design of a synthetic microorganism from the ground up might be a reasonable experimental agenda for a microbial biofuel project, other questions are more natural in human medicine.

### Human-Specific Problems

It became fairly clear by the height of the era of evidence-based medicine that prevention has the highest ratio of benefit versus cost to society. Prevention is most effective when coupled to prediction. Interventions have risks and costs, so should be targeted to those predicted to have sufficient risk to benefit from a preventive intervention. Therefore a major goal of systems biology is to develop personalized predictive risk assessments. These risk assessments can be absolute (e.g., an individual’s risk at the time of birth or at the time of a clinic visit) or conditional (e.g., risk if the individual stopped smoking). The types of outcomes that are useful to predict for an individual include whether or not a vaccination is protective, and whether a disease might occur and when it is likely to do so. A particular world health challenge today is the unpredictability of the effectiveness of the BCG vaccine for tuberculosis. Individuals who receive the vaccine have little way of knowing whether or not they will develop protective immunity [3]. The identification of biomarkers enabling such a prediction would be useful for counseling individuals as well as developing society-wide measures for preventing tuberculosis. Predictions at the level of a cell or tissue include what developmental programs these cells might follow or specific outcomes such as apoptosis, endocytosis, or mobility.

Other types of questions that can be asked in human systems include:

1. The identification of key information-transducing nodes in a system, as well as the inference of nodes that must be part of the system but are currently unsuspected or unknown. Such inference would be based on experimental results that differed from predicted results.
2. The identification of therapeutic drug targets [19].

3. Prediction of the effects of a gene knockout or knock-down, or overexpression of that gene.
4. Most generally, prediction of the effects of any perturbation to a system, including chemical and environmental perturbations.
5. Prediction of a particular property of a node or edge, such as Gene Ontology category, the presence of negative regulation, or the presence of positive or negative feedback.
6. Prediction of the concentration or flux or production of a specific gene, mRNA, or protein under a specific set of conditions at a specific time in a system.

The most successful projects will define exactly what questions will be addressed by the project before the project begins, and will define measurable quantitative variables that will be predicted by models [23]. The ability of a research group to precisely and accurately make quantitative predictions is one useful measure of whether a particular research methodology and approach is working effectively. Some questions lend themselves more naturally to quantitative predictions, such as those related to flux or concentration of small molecules. Others, such as the identification of drug targets, are very difficult to evaluate on any timescale less than a few decades.

Systems biology must be applied to all human diseases, as the urgency of clinical discovery is paramount. However, from the point of view of leveraging human systems biology research to gain fundamental understanding of a system, some human diseases may lend themselves more readily to knowledge discovery. These will be the diseases for which the etiology lies contained in a fairly small network and can be understood by understanding the interactions of a relatively few nodes. Thus, one might predict that research with Noonan Syndrome, which appears to be related to defects within the Ras signaling network may lead more immediately to fundamental understanding than research in type 1 diabetes, which appears to have etiologic effects spread throughout many cell types that arise from both genetic and environmental influences, and that act at disparate timepoints throughout the life of the individual.

## Examples

Research in Noonan Syndrome represents an elegant example of systems biology as applied to human disease. Phenotypes of Noonan Syndrome include congenital heart malformation, short stature, and learning problems. Approximately 1 in 2,000 children are born with Noonan Syndrome. The range and severity of features can vary greatly between individuals, suggesting that there are mul-

tiiple causes of the syndrome that affect common etiologic pathways, but in subtly different manners. Frequent transmission from parent to child suggested a genetic defect with autosomal dominant inheritance. Traditional genetic approaches mapped Noonan Syndrome to chromosome 12q24. By 2001, it was shown that about half of all individuals with Noonan syndrome were caused by a mutation of the PTPN11 gene at position 12q24. The PTPN11 gene encodes a tyrosine phosphatase. This discovery provided the first insight into the formation of a predictive model for Noonan Syndrome. With an underlying hypothesis that Noonan Syndrome was the emergent result of the output of an unknown dysfunctional network, the discovery that PTPN11 was part of that network allowed orthologous data sets and analysis to be leveraged [28,29]. PTPN11 is a node in an intracellular signaling network, including the traditional Ras/MAP kinase signaling pathway. This network governs cell division and differentiation. This network was fairly well characterized prior to the discovery of the association of PTP11 with Noonan Syndrome, so the network knowledge could be immediately applied to predict other nodes that, when perturbed, might also result in a Noonan Syndrome phenotype [13]. When tested, using an iterative network-refinement approach, several of these genes were also found to be fundamental to Noonan Syndrome etiology. These genes include RAF1, SOS1, HRAS, KRAS, and RAF1 [36,39,44,46,52].

Genetic association studies show great promise for breaking into etiologic network models for other human diseases. The technology for association studies for autoimmune diseases has been progressing steadily since the 1970s. However, until 2006, the only major genetic association with these diseases had been the MHC locus, which encodes the molecular complexes that train the immune system to recognize particular peptides as activators. The accumulated progress in technology moved research across a discovery threshold early in the twenty-first century and a large number of additional gene associations in a variety of autoimmune diseases have been discovered, as well in other complex human diseases. These include the discovery of IFIH1 for type 2 diabetes [45], ITPR3 for type 1 diabetes [40], and IL7R for multiple sclerosis [15,22]. Further progress in these diseases will be somewhat slower than for Noonan Syndrome, but there is every reason to believe that these larger and more complex networks can be modeled in clinically useful ways by applying the techniques discussed in the previous section.

In addition to genetic association studies, data analysis from high-throughput sources can also help enumerate key nodes in human networks. Carefully curated analyzes of orthologous high-throughput data sets can yield lists of

nodes that are likely to be fundamental to predictive network models. One of the more promising human systems for this systems biology approach is modeling macrophage activation through innate immune receptors [2]. Non-predictive network diagrams based on extensive literature analysis are available for the macrophage [34]. Such diagrams facilitate the development of predictive network hypotheses [35].

Rather than breaking into an unknown network, Bergholdt et al. [7] have approached the type 1 diabetes network by assuming that it is a subnetwork of current iterations of the global human protein-protein interaction network. They take key-node analysis a step further by developing a statistic for key subnetworks. Significance is computed from genetic association data for all genes in the subnetwork, as well as the interactions between these genes. This approach is amenable to further refinement by the incorporation of additional orthologous data sets, such as function and expression, into the significance metric. Bergholdt et al. succeed in identifying a number of candidate subnetworks that may transduce signals that cause diabetes.

Human macrophages are tractable experimental systems even when taken out of the context of the human body, and much of the macrophage regulatory network operates in a near native fashion even when isolated from other human cells. Because systems biology of the human immune system can in cases be reduced to the analysis of single cells, it shares some simplicities with prokaryotic systems biology. Systems biology of the human immune system is therefore a natural initial focus for human systems biology.

A large number of high-throughput data sets have been acquired from macrophages, particularly time-series following stimulation through toll-like receptors. Several preliminary comparative network analyzes are available [21,37]. Reasonable near-term expectations from such studies, which may involve dozens of researchers and years of research, are the addition of one or several nodes and key interactions to a basic network such as the identification of the role of ATF3 as a important negative regulator of macrophage activation [14] or the bulk identification of many key nodes without their interactions such as the identification of all transcriptionally active transcription factors [33,41]. Studies of signaling networks in T-cells have also resulted in predictive models, but are impeded by the resistance of T-cells to some forms of genetic perturbations, such as RNAi interference [25].

A lot of clinical value is currently being generated through analyzes of high-throughput data sets that result in predictive models for disease [27]. In some cases these

analyzes are done on only one type of data, such as microarray data, and are not orthologous data sets in the truest sense. Furthermore, the results are frequently statements of correlation, rather than causation, and thus are a fairly preliminary step towards a true understanding of the system, or a true predictive etiologic model. Nevertheless these predictions are amenable to iterative refinement with a good prospect that such refinements will eventually elevate these models to the level of understanding. Examples of such efforts include diagnosis and stratification.

Statistical methodologies such as discriminant or principle components analysis can extract variables that are highly informative for classifying systems. Quantitative values can be obtained from many thousand nodes in systems with known properties. For example, one can measure the expression of thousands of genes in a set of malignant tumors and controls. Simplified classifiers can be extracted from these many thousand variables. Clinically useful stratification of breast cancers into subtypes that respond preferentially to different therapeutic interventions has been one example of this approach [16,49,50]. These multivariate approaches also show promise for diagnosing disease, including acute infections [38], arthritis [4], and diabetes [24]. These diagnostic techniques show promise for detecting presymptomatic disease, in which cases the semantic line between prediction and diagnosis blurs considerably. These methodologies can also stratify disease, allowing for the tailoring of specific therapies to the disease. Additionally, responses to diseases, medications, therapies, or the environment can themselves be stratified, permitting the early identification of adverse reactions and changes towards a more positive therapeutic direction.

## Future Directions

The most productive element of the systems biology paradigm is collaboration. Collaboration between scientists with similar skills allows projects of greater size and complexity to be tackled. Collaboration between scientists with different knowledge and skill sets permits knowledge to be combined in manners previously unimaginable and opens previously inaccessible vistas of knowledge. For decades, both types of collaboration have been widely used in other branches of science, such as meteorology, astronomy, and particle physics. Therefore the most promising future directions for system biology are those that maximize collaborations.

These collaborations will be enabled by new methods of data sharing, most likely leveraging the worldwide web. Interactive and updateable sources of information such as

wikis and blogs will be key. Many online scientific journals have recently begun to encourage online commenting of papers. New methods of collaboration may evolve. These may involve virtual collaboration and might leverage peer-to-peer networks such as Facebook. The Nature publishing group has initiated such a tool, Nature Network.

New societal methods of career recognition and funding will likely stimulate systems biology. Traditional methods have recognized single-author or first author papers and tend to focus on journal publishing as a measure of productivity and utility. New methods will recognize collaboration and non-traditional measures of productivity, such as the creation of online resources. Recently, the NIH has altered its tenure track rules in recognition of these needs. Many of these changes will require a new generation of researchers willing to embrace them. Progress in systems biology will be increasingly driven by distributed worldwide collaborations that may involve researchers unaware of each other, with the collaborative element being contributions to virtual online projects. Projects will involve both specialists and generalists. In many cases data will be generated by a specialist in an experimental technique, and analyzed by distinct specialists in analytical techniques. Generalists will need to have skills in multiple disciplines and will be needed to direct projects and integrate data. Among the traditional twentieth-century fields of biology, immunology required extensive multidisciplinary knowledge from its practitioners; it is not surprising that many of the pioneers of systems biology were immunologists. Investigators trained in the clinical practice of medicine will be needed to set clinical priorities and recognize opportunities for the transfer of information from bench to bedside. Walls between disciplines will need to erode.

New and improved technologies will increasingly yield lower-costs, higher-throughputs, better sensitivity, better reproducibility and precision, better accuracy, and more comprehensivity [17,20]. Sequencing technologies continue to show promise for all of these characteristics, and dropping costs may permit them to underlie many measurements. Transcriptomics is currently performed mostly with arrays which may have fundamental upper limits on accuracy. Future transcriptomics may breach the accuracy barrier by employing transcript enumeration technologies that aim to sequence, for example, every mRNA in a single cell. Improvement in computational prediction and direct measurement of transcription factor binding to targets will permit these interactions to be confidently added to networks, and will vastly improve the predictive power of models of nuclear regulation. New methods of creating, storing, and sharing medical information electroni-

cally must be adopted that allow efficient use of aggregate information for research while preserving patient confidentiality and their rights to control their own information.

Effective improvements to human health cannot be made solely by a small group of individuals. Society must participate. Individuals must be informed, must support research, and must be willing to participate in and receive therapeutic interventions. Systems biology must therefore be alert to the broad needs of society [10]. These needs can help focus priorities. A society willing to modify its diet but reluctant to take pharmaceuticals would inform systems biology to prioritize models that predict the effects of altered diets over models of drug metabolism, and vice versa. Because human systems biology relies on human participation in research, such as in the Framingham Study and in gene association studies, efforts must be made to continue to educate all people in a deep inquiry-based scientific knowledge and to encourage them to share their personal medical histories and experience with the common knowledge pool. Current cultural shifts in the US towards increased guarding of personal information and mistrust of scientific information will create substantial challenges for the next generation of systems biologists.

## Bibliography

### Primary Literature

1. Aderem A (2007) Systems biology. *Curr Opin Biotechnol* 18(4):331–332
2. Aderem A, Hood L (2001) Immunology in the post-genomic era. *Nat Immunol* 2(5):373–5
3. Aderem A, Smith KD (2004) A systems approach to dissecting immunity and inflammation. *Semin Immunol* 16(1):55–67
4. Allantaz F, Chaussabel D, Stichweh D, Bennett L, Allman W, Mejias A, Ardura M, Chung W, Wise C, Palucka K, Ramilo O, Punaro M, Banchereau J, Pascual V (2007) Blood leukocyte microarrays to diagnose systemic onset juvenile idiopathic arthritis and follow the response to IL-1 blockade. *J Exp Med* 204(9):2131–44
5. Auffray C, Imbeaud S, Roux-Rouquie M, Hood L (2003) From functional genomics to systems biology: concepts and practices. *C R Biol* 326(10–11):879–92
6. Basso K, Margolin AA, Stolovitzky G, Klein U, Dalla-Favera R, Califano A (2005) Reverse engineering of regulatory networks in human B cells. *Nat Genet* 37(4):382–90
7. Bergholdt R, Størling ZM, Lage K, Karlberg EO, Olason PL, Aalund M, Nerup J, Brunak S, Workman CT, Pociot F (2007) Integrative analysis for finding genes and networks involved in diabetes and other complex diseases. *Genome Biol* 8(11):R253
8. Beyer A, Bandyopadhyay S, Ideker T (2007) Integrating physical and genetic maps: from genomes to interaction networks. *Nat Rev Genet* 8(9):699–710
9. Bornholdt S (2005) Systems biology. Less is more in modeling large genetic networks. *Science* 310(5747):449–51



10. Buchanan A, Califano A, Kahn J, McPherson E, Robertson J, Brody B (2002) Pharmacogenetics: ethical issues and policy options. *Kennedy Inst Ethics J* 12(1):1–15
11. Chaussabel D (2004) Biomedical literature mining: challenges and solutions in the 'omics' era. *Am J Pharmacogenomics* 4(6):383–93
12. Daly AK (2007) Individualized drug therapy. *Curr Opin Drug Discov Devel* 10(1):29–36
13. Gelb BD, Tartaglia M (2006) Noonan syndrome and related disorders: dysregulated RAS-mitogen activated protein kinase signal transduction. *Hum Mol Genet* 15, Spec No 2:R220–6
14. Gilchrist M, Thorsson V, Li B, Rust AG, Korb M, Roach JC, Kennedy K, Hai T, Bolouri H, Aderem A (2006) Systems biology approaches identify ATF3 as a negative regulator of Toll-like receptor 4. *Nature* 441(7090):173–8
15. Gregory SG, Schmidt S, Seth P, Oksenberg JR, Hart J, Prokop A, Caillier SJ, Ban M, Goris A, Barcellos LF, Lincoln R, McCauley JL, Sawcer SJ, Compston DA, Dubois B, Hauser SL, Garcia-Blanco MA, Pericak-Vance MA, Haines JL; for the Multiple Sclerosis Genetics Group (2007) Interleukin 7 receptor alpha chain (IL7R) shows allelic and functional association with multiple sclerosis. *Nat Genet* 39(9):1083–91
16. Hoadley KA, Weigman VJ, Fan C, Sawyer LR, He X, Troester MA, Sartor CI, Rieger-House T, Bernard PS, Carey LA, Perou CM (2007) EGFR associated expression profiles vary with breast tumor subtype. *BMC Genomics* 8(1):258
17. Hood L (2002) A personal view of molecular technology and how it has changed biology. *J Proteome Res* 1(5):399–409
18. Hood L, Galas D (2003) The digital code of DNA. *Nature* 421(6921):444–8
19. Hood L, Perlmutter RM (2004) The impact of systems approaches on biological problems in drug discovery. *Nat Biotechnol* 22(10):1215–7
20. Hood L, Heath JR, Phelps ME, Lin B (2004) Systems biology and new technologies enable predictive and preventative medicine. *Science* 306(5696):640–3
21. Iliev DB, Roach JC, Mackenzie S, Planas JV, Goetz FW (2005) Endotoxin recognition: In fish or not in fish? *FEBS Lett* 579(29):6519–6528
22. International Multiple Sclerosis Genetics Consortium, Hafler DA, Compston A, Sawcer S, Lander ES, Daly MJ, De Jager PL, de Bakker PI, Gabriel SB, Mirel DB, Iverson AJ, Pericak-Vance MA, Gregory SG, Rioux JD, McCauley JL, Haines JL, Barcellos LF, Cree B, Oksenberg JR, Hauser SL (2007) Risk alleles for multiple sclerosis identified by a genomewide study. *N Engl J Med* 357(9):851–62
23. Janes KA, Yaffe MB (2006) Data-driven modelling of signal-transduction networks. *Nat Rev Mol Cell Biol* 7(11):820–8
24. Kaizer EC, Glaser CL, Chaussabel D, Banchereau J, Pascual V, White PC (2007) Gene expression in peripheral blood mononuclear cells from children with diabetes. *J Clin Endocrinol Metab* 92(9):3705–11
25. Kemp ML, Wille L, Lewis CL, Nicholson LB, Lauffenburger DA (2007) Quantitative network signal combinations downstream of TCR activation can predict IL-2 production response. *J Immunol* 178(8):4984–92
26. Kitano H (2002) Computational systems biology. *Nature* 420(6912):206–10
27. Kitano H (2007) The theory of biological robustness and its implication in cancer. *Ernst Schering Res Found Workshop*. (61):69–88
28. Kratz CP, Schubert S, Bollag G, Niemeyer CM, Shannon KM, Zenker M (2006) Germline mutations in components of the Ras signaling pathway in Noonan syndrome and related disorders. *Cell Cycle* 5(15):1607–11
29. Lee ST, Ki CS, Lee HJ (2007) Mutation analysis of the genes involved in the Ras-mitogen-activated protein kinase (MAPK) pathway in Korean patients with Noonan syndrome. *Clin Genet* 72(2):150–5
30. Ma'ayan A, Jenkins SL, Neves S, Hasseldine A, Grace E, Dubin-Thaler B, Eungdamrong NJ, Weng G, Ram PT, Rice JJ, Kershensbaum A, Stolovitzky GA, Blitzer RD, Iyengar R (2005) Formation of regulatory patterns during signal propagation in a mammalian cellular network. *Science* 309(5737):1078–83
31. Ma'ayan A, Jenkins SL, Goldfarb J, Iyengar R (2007) Network analysis of FDA approved drugs and their targets. *Mt Sinai J Med* 74(1):27–32. Ma'ayan A, Blitzer RD, Iyengar R (2005) Toward predictive models of mammalian cells. *Annu Rev Biophys Biomol Struct* 34:319–49
32. Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, Califano A (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics* 7 Suppl 1:S7
33. Natarajan M, Lin KM, Hsueh RC, Sternweis PC, Ranganathan R (2006) A global analysis of cross-talk in a mammalian cellular signalling network. *Nat Cell Biol* 8(6):571–80
34. Oda K, Kitano H (2006) A comprehensive map of the toll-like receptor signaling network. *Mol Syst Biol* 2:2006.0015
35. Oda K, Matsuoka Y, Funahashi A, Kitano H (2005) A comprehensive pathway map of epidermal growth factor receptor signaling. *Mol Syst Biol* 1:2005.0010
36. Pandit B, Sarkozy A, Pennacchio LA, Carta C, Oishi K, Martinelli S, Pogna EA, Schackwitz W, Ustaszewska A, Landstrom A, Bos JM, Ommen SR, Esposito G, Lepri F, Faul C, Mundel P, Lopez Siguero JP, Tenconi R, Selicorni A, Rossi C, Mazzanti L, Torrente I, Marino B, Digilio MC, Zampino G, Ackerman MJ, Dallapiccola B, Tartaglia M, Gelb BD (2007) Gain-of-function RAF1 mutations cause Noonan and LEOPARD syndromes with hypertrophic cardiomyopathy. *Nat Genet* 39(8):1007–12
37. Purcell MK, Smith KD, Hood L, Winton JR, Roach JC (2006) Conservation of toll-like receptor signaling pathways in teleost fish. *Comp Biochem Physiol Part D Genomics Proteomics* 1(1):77–88
38. Ramilo O, Allman W, Chung W, Mejias A, Ardura M, Glaser C, Wittkowski KM, Piqueras B, Banchereau J, Palucka AK, Chaussabel D (2007) Gene expression patterns in blood leukocytes discriminate patients with acute infections. *Blood* 109(5):2066–77
39. Razzaque MA, Nishizawa T, Komoike Y, Yagi H, Furutani M, Amo R, Kamisago M, Momma K, Katayama H, Nakagawa M, Fujiwara Y, Matsushima M, Mizuno K, Tokuyama M, Hirota H, Muneuchi J, Higashinakagawa T, Matsuoka R (2007) Germline gain-of-function mutations in RAF1 cause Noonan syndrome. *Nat Genet* 39(8):1013–7
40. Roach JC, Deutsch K, Li S, Siegel AF, Bekris LM, Einhaus DC, Sheridan CM, Glusman G, Hood L, Lernmark AA, Janer M on behalf of the Swedish Childhood Diabetes Study Group and the Diabetes Incidence in Sweden Study Group. (2006) Genetic mapping at 3-kilobase resolution reveals inositol 1,4,5-triphosphate receptor 3 as a risk factor for type 1 diabetes in Sweden. *Am J Hum Genet* 79(4):614–627



41. Roach JC, Smith KD, Strobe KL, Nissen SM, Haudenschild CD, Zhou D, Vasicek TJ, Held GA, Stolovitzky GA, Hood L, Aderem A (2007) Transcription factor expression in lipopolysaccharide-activated peripheral blood derived mononuclear cells. *PNAS* 104(41):16245–50
42. Roeder RG (2003) The eukaryotic transcriptional machinery: complexities and mechanisms unforeseen. *Nat Med* 9(10):1239–44
43. Schlitt T, Brazma A (2007) Current approaches to gene regulatory network modelling. *BMC Bioinformatics*. 8 Suppl 6:S9
44. Schubbert S, Zenker M, Rowe SL, Boll S, Klein C, Bollag G, van der Burgt I, Musante L, Kalscheuer V, Wehner LE, Nguyen H, West B, Zhang KY, Sistermans E, Rauch A, Niemeyer CM, Shannon K, Kratz CP (2006) Germline KRAS mutations cause Noonan syndrome. *Nat Genet* 38(3):331–6
45. Smyth DJ, Cooper JD, Bailey R, Field S, Burren O, Smink LJ, Guja C, Ionescu-Tirgoviste C, Widmer B, Dunger DB, Savage DA, Walker NM, Clayton DG, Todd JA (2006) A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nat Genet* 38:617–9
46. Sovik O, Schubbert S, Houge G, Steine SJ, Norgard G, Engelsen B, Njolstad PR, Shannon K, Molven A (2007) De novo HRAS and KRAS mutations in two siblings with short stature and neurocardio-facio-cutaneous features. *J Med Genet* 44(7):e84
47. Tian Q, Stepaniants SB, Mao M, Weng L, Feetham MC, Doyle MJ, Yi EC, Dai H, Thorsson V, Eng J, Goodlett D, Berger JP, Gunter B, Linsley PS, Stoughton RB, Aebersold R, Collins SJ, Hanlon WA, Hood LE (2004) Integrated genomic and proteomic analyses of gene expression in Mammalian cells. *Mol Cell Proteomics* 3(10):960–9
48. Tijan R (1995) Molecular machines that control genes. *Scientific American*. 272(2):54–61
49. van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415(6871):530–6
50. Weigelt B, Hu Z, He X, Livasy C, Carey LA, Ewend MG, Glas AM, Perou CM, Van't Veer LJ (2005) Molecular portraits and 70-gene prognosis signature are preserved throughout the metastatic process of breast cancer. *Cancer Res* 65(20):9155–8
51. Weston AD, Hood L (2004) Systems biology, proteomics, and the future of health care: toward predictive, preventative, and personalized medicine. *J Proteome Res* 3(2):179–96
52. Zenker M, Horn D, Wiczorek D, Allanson J, Pauli S, van der Burgt I, Doerr HG, Gaspar H, Hofbeck M, Gillesen-Kaesbach G, Koch A, Meinecke P, Nowak A, Rauch A, Reif S, von Schnakenburg C, Seidel H, Wehner LE, Zweier C, Bauhuber S, Matejas V, Kratz CP, Thomas C, Kutsche K (2007) SOS1 is the second most common Noonan gene but plays no major role in cardio-facio-cutaneous syndrome. *J Med Genet* 44(10):651–6
- Carroll SB (2004) From DNA to diversity: Molecular genetics and the evolution of animal design. Blackwell Publishing Limited, Boston
- Davidson EH (2006) The regulatory genome: Gene regulatory networks in development and evolution. Academic Press, San Diego
- Fall CP et al (2002) Computational cell biology. Springer, New York
- Goodsell DS (1992) The machinery of life. Springer, New York
- Kaneko K (2006) Life: An introduction to complex systems biology. Springer, New York
- Kauffman SA (1993) The origins of order: Self-organization and selection in evolution. Oxford University Press, Oxford
- Kitano H (2001) Foundations of systems biology. MIT Press, Cambridge
- Latchman DS (2004) Eukaryotic transcription factors. Elsevier, Amsterdam
- Michal G (1999) Biochemical pathways. Wiley, New York
- Palsson BO (2006) Systems biology: Properties of reconstructed networks. Cambridge University Press, Cambridge
- Voit EO (2000) Computational analysis of biochemical systems: A practical guide for biochemists and molecular biologists. Cambridge University Press, Cambridge

## Systems Biology, Introduction to

ROBERT A. MEYERS

Ramtech Limited, Larkspur, USA

Systems biology is a relatively new branch of life sciences that is highly interdisciplinary and aims to provide a quantitative analysis and understanding of life systems within a mathematical framework. Systems biology is the study of the interconnected aspect of molecular, cellular, tissue, whole animal and ecological processes, and comprises mathematical and mechanistic studies of dynamical, mesoscopic, open, spatiotemporally defined, nonlinear, complex systems that are far from thermodynamic equilibrium. The animating principle of systems biology is that the most important behaviors and networks of interacting elements (e.g., molecules, cells, etc.), rather than by individual elements and interactions. This really necessitates the integrated application of high-throughput measurement technologies and advanced computational methods to model and predict biological responses. The 11 articles comprising this section describe the basic aspects and processes of systems biology and the relationship of systems biology to complexity and systems science. The basis and content of each of the 11 articles in this section are described below. There are six additional articles that were nominated and accepted into other sections which are connected to complexity in Systems Biology, and supplement the 11 articles in this section. These six are listed at the end of this Introduction.

### Books and Reviews

- Alon U (2006) An introduction to systems biology: Design principles of biological circuits. Chapman and Hall, Norwell
- Bower JM, Bolouri H (2001) Computational modeling of genetic and biochemical networks. MIT Press, Cambridge
- Capra F (1996) The web of life. Doubleday, New York

With rare exceptions, all known living organisms encode their genetic material in the form of double-stranded DNA, in one or more chromosomes, collectively referred to as the “genome” (► [Genome Organization](#)). The genome includes most of the information needed by the cells to stay alive, to differentiate into new cell types, and to perform their functions in the context of the organism. As such, it is the ultimate resource for identifying the full set of components in the living system. Eukaryotic genomes are much larger than strictly needed to encode the relatively modest set of genes in them, but several mechanisms give rise to a very complex transcriptome. It is necessary to understand how the genome is organized and how it evolved and be able to build suitable null hypotheses for testing whether predictions are likely to be real, and thereby understand whether specific observations are likely to be biologically meaningful.

Sequence data alone is currently of limited use for identifying the functional elements of a genome and for elucidating how these elements interact to control physiological processes.

The field of functional genomics aims to meet these challenges using the sequence data as a blueprint (► [Functional Genomics for Characterization of Genome Sequences](#)). In a broad sense, functional genomics is defined as the large-scale experimental study of gene function and interactions. In the above article, the techniques and challenges in functional genomics are illustrated in the context of an ever more common case scenario: given a complete genome sequence, how does one figure out what the sequence means? The current state of the art is interpreted to arrive at the function or functions of a given genome sequence.

Among the most fundamental problems in biology is deciphering the relationship between genotype and phenotype in the complex system of life forms. Understanding a complex system broadly requires that one (1) identify the elements, (2) determine the function of each element, (3) identify and characterize the interactions between elements, and (4) assemble all of this information into a mathematical model that accurately simulates the system and predicts its responses to novel perturbations (► [Systems Genetics and Complex Traits](#)). In the context of systems biology, this process begins with sequencing an organism’s genome and identifying the functional elements, e. g. the genes. There are numerous methods for determining the function(s) of a given gene product, where “function” can describe both the specific biochemical activity carried out by the gene product and the role that activity plays in the organism’s response to an environmental or developmental stimulus. Understanding com-

plex genetics is of increasing relevance to the study of human health and is essential to the development of predictive, preventive, and personalized medicine.

Systems biology is the derivation of emergent properties of a multicomponent biological system through the construction of quantitative predictive models. These predictive models are typically formulated as networks. Human systems biology is the practice of systems biology to elucidate emergent properties of humans or any subcomponent of the human body, such as a cell (► [Systems Biology of Human Immunity and Disease](#)). Human systems present challenges and opportunities not found in other organisms. Most problems in human systems biology are motivated by a desire to understand, predict, prevent, ameliorate, or cure a human disease. The human immune system is a mediator for most human diseases, and is the key system mediating autoimmune and infectious diseases.

Understanding the operation of cellular networks is probably one of the most challenging and intellectually exciting scientific fields today (► [Biological Models of Molecular Network Dynamics](#)). Cellular networks are some of the most complex natural systems we know. Even in a “simple” organism such as *E. coli*, there are at least four thousand genes with many thousands of interactions between molecules of many different sizes. With the availability of new experimental and theoretical techniques, our understanding of the operation of cellular networks has made great strides in the last few decades. An important outcome of this work is the development of predictive quantitative models. Such models of cellular function will have a profound impact on our ability to manipulate living systems which will lead to new opportunities for generating energy, mitigating our impact on the biosphere and last but not least, opening up new approaches and understanding of important disease states such as cancer and aging.

Biological research over the past century or so has been dominated by reductionism – identifying and characterizing individual biomolecules – and has enjoyed enormous success. Throughout this history, however, it has become increasingly clear that an individual biomolecule can rarely account for a discrete biological function on its own. A biological process is almost always the result of a complex interplay of relationships amongst biomolecules, and the treatment of these relationships as a graph is a natural and useful abstraction (► [Biomolecular Network Structure and Function](#)).

Broadly speaking, a biomolecular network is a graph representation of relationships (of which there are many types) amongst a group of biomolecules. Vertices or nodes

represent biomolecules, including macromolecules such as genes, proteins, and RNAs, or small biomolecules like amino acids, sugars, and nucleic acids. An edge or link between two vertices indicates a relationship between the corresponding biomolecules, which could include physical interaction, genetic interaction, or a regulatory relationship (e. g., the protein product of gene A regulates the expression of gene B). This abstraction, although simplifying, converts a complex web of biological relationships into a mathematical graph, from which we can study its structural features as well as their implications on biological functions.

Data integration and model building have become essential activities in biological research as technological advancements continue to empower the measurement of biological data of increasing diversity and scale (► [Biological Data Integration and Model Building](#)). High-throughput technologies provide a wealth of global data sets (e. g., genomics, transcriptomics, proteomics, metabolomics), and the challenge becomes how to integrate this data to maximize the amount of useful biological information that can be extracted. Integrating biological data is important and challenging because of the nature of biology. Biological systems have evolved over the course of billions of years, and in that time biological mechanisms have become very diverse, with molecular machines of intricate detail. Thus, while there are certainly great general scientific principles to be distilled – such as the foundational evolutionary theory – much of biology is found in the details of these evolved systems. This emphasis on the details of systems and the history by which they came into being (i. e. evolution) are distinct features of biology as a science, and influence the need for large-scale data integration. Also, biological systems are responsive to varying environments, with potential system states influenced by the combinatorics of all possible molecular and environmental perturbations. Thus, data space in this realm is extraordinarily large. There is no shortage of possibilities to explore, and vast amounts of data will be needed to reengineer biological systems and understand their workings at a (near) complete level of detail or at a level where accurate prediction can be achieved.

Another reason data integration is essential, is that biology arises through the complex interworking of many components. Thus, to understand biology we must integrate biological data in a way that will allow us, not only to access all our acquired data efficiently, but even more importantly allow us to study, predict, modify, and even engineer the emergent properties of biological systems, where the whole is more than the sum of the parts.

Boolean Networks are a class of discrete dynamical systems that can be characterized by the interaction of a set of Boolean variables. Random Boolean Networks, which are ensembles of random network structures, are a simple model class for studying dynamical properties of gene regulatory networks (► [Boolean Modeling of Biological Networks](#)). Boolean Networks have been used as generic models for dynamics of complex systems of interacting entities, such as social and economic networks, neural networks, as well as gene and protein interaction networks. Despite their conceptual simplicity, Boolean Networks exhibit complex nonlinear behaviors that are, to this day, a challenging object of investigation for theoretical physicists and mathematicians. Further, a discretization of gene expression is often regarded as an experimentally justifiable simplification making Boolean network models attractive tools for the biologist.

Many processes in cell biology, such as those that carry out metabolism, the cell cycle, and various types of signaling, are comprised of biochemical reaction networks. It has proven useful to study these networks using computer simulations because they allow us to quantitatively investigate hypotheses about the networks. Deterministic simulations are sufficient to predict average behaviors at the population level, but they cannot address questions about noise, random switching between stable states of the system, or the behaviors of systems with very few molecules of key species. These topics are investigated with stochastic simulations (► [Stochastic Models of Biological Processes](#)). The dominant types of stochastic simulation methods that are used to investigate biochemical reaction networks are covered, as well as some of the results that have been found with them. As new biological experiments continue to reveal more detail about biological systems, and as computers continue to become more powerful, researchers will increasingly turn to simulation methods that can address stochastic and spatial details.

Systems biology has various definitions. Common features among accepted definitions generally involve the description and analysis of interacting biomolecular components. Systems analysis of a biological network is quickly demonstrating its utility as it helps to characterize biomolecular behavior that could not otherwise be produced by the individual components alone. Three areas in which systems analysis has been implemented in biology include: (1) the generation and statistical analysis of high-throughput data in an effort to catalog and characterize cellular components; (2) the construction and analysis of computational models for various biological systems (e. g., metabolism, signaling, and transcriptional regulation); and (3) the integration of the knowledge of parts and

computational models to predict and engineer biological systems (synthetic biology). Metabolism, as a system, has played an important role in the development of systems biology, especially in the modeling sense (► [Metabolic Systems Biology](#)). This is because the network components (e. g., enzymes and metabolites) have been studied in detail for decades, and many links between components have been experimentally characterized. Metabolic systems biology, compared to systems biology in general, entails the computational analysis of these enzymes and metabolites and the metabolic pathways in which they participate. Metabolic systems biology, using genome-scale metabolic network reconstructions and their models, has helped (1) to elucidate biomolecular function; (2) to predict phenotypic behavior; (3) to discover new biological knowledge; and (4) to design experiments for engineering applications. Constraint-based methods have played a pivotal role in the analysis of large and genome-scale metabolic networks. The structure, mathematical formulation, and analytical techniques of constraints-based methods have also paved the way for the successful modeling of other complex biological networks, such as transcriptional regulation and signaling networks.

Finally, ecological systems are paradigmatic examples of complex systems. For example, consider the thousands of species interacting in complex ways within rich communities such as tropical rainforests or coral reefs. The most pressing questions ecologists face deal with concepts such as stability, resilience, thresholds and non-linearities which are at the core of the sciences of complexity. How robust are these cathedrals of biodiversity? At which rate will they disassemble as a consequence of global change. For example, one of the long-standing questions in ecology is the relationship between complexity and stability. A review of some of the applications of the complexity sciences in the realm of ecological systems is presented (► [Ecological Systems](#)). Predicting the consequences of global change on biodiversity requires an interdisciplinary approach in which complexity approaches may be very useful. Information theory and diversity, networks, complex dynamics and spatiotemporal dynamics are all discussed relative to ecology as a complex system.

Additional articles connected to Systems Biology:

- [Consciousness and Complexity](#)
- [Molecular Evolution, Networks in](#)
- [Complexity in Systems Level Biology and Genetics: Statistical Perspectives](#)
- [Fractals in Biology](#)
- [Biochemistry, Chaotic Dynamics, Noise, and Fractal Space in](#)
- [Exobiology and Complexity](#)

## Systems and Control, Introduction to

MATTHIAS KAWSKI

Department of Mathematics and Statistics,  
Arizona State University, Tempe, USA

Control of dynamical systems has a long history: Watt's automatic centrifugal governor designed to regulate steam engines in the 1780s is considered a precursor of modern feedback control. Similarly, Bernoulli's work in the 1690s on the brachistochrone problem is a progenitor of optimal control. A distinguishing external feature of *controlled* dynamical systems is the presence of inputs that interact with the system in the form of deliberate controls, or unavoidable perturbations. Commonly control systems also have outputs (observations) that typically provide only incomplete information about the state of the system. Systems and control theory utilizes a broad array of mathematical disciplines – but a uniting, characteristic feature is the kind of questions being asked. In the 1950s the field became more formalized, it began to develop its own distinctive identity, and it has rapidly evolved ever since. Naturally, the linear theory developed first and quickly became a mature subject, with controllers now pervading everyday life, from thermostats, dozens of controllers in every automobile, to cell phones, attitude control of satellites, electric power networks, highway traffic control, to name just a few.

The articles in this section focus on the more recent developments, articulating how the fundamental problems and questions of systems and control theory are addressed in ever new settings, on new tools, and new applications.

Whereas in classical dynamical systems one tries to predict the uniquely determined future, a basic distinguishing question in control theory asks whether it is possible to steer the system to any desired target. This question about *controllability* has a simple and elegant answer in the case of linear systems. While much also is known for nonlinear systems, many challenges remain in both the finite dimensional setting [see ► [Finite Dimensional Controllability](#)] and in the setting of *distributed systems* modeled by partial differential equations [see ► [Control of Non-linear Partial Differential Equations](#)].

Assuming controllability using controls that are functions of time, the next question asks whether one can *automate* the control using feedback, that is, by making it a function of the state or of a measured output. Whereas for linear systems the theory is straightforward, unavoidable topological obstacles make nonlinear feedback stabilization problems much more challenging [see ► [Stability and Feedback Stabilization](#)]. A notion dual to control-



lability is that of observability: Rather than asking about the map from inputs to the state, a system is observable if the history of the outputs (measurements) determines the state of the system [see ► [Observability \(Deterministic Systems\) and Realization Theory](#)]. The typical engineering problem involves much more complex multi-tasking, dealing with multiple objectives in the presence of diverse constraints. Much of this well developed theory of output regulation, disturbance rejection, while attending to performance criteria, has recently been generalized to nonlinear systems [see ► [System Regulation and Design, Geometric and Algebraic Methods in](#)]. Generally, if there is one, then there are many control strategies that achieve a certain task – naturally one likes to single out a strategy that is the best. This leads to optimal control theory, which supersedes much of the classical calculus of variations, and which has rich interfaces with Lagrangian and Hamiltonian mechanics [see ► [Maximum Principle in Optimal Control](#)].

In many applications all one has to work with are sampled data for inputs and measured outputs. Moreover, commonly these are also corrupted by measurement errors and noises. A fundamental problem is to devise algorithms to *identify* the best model system that could generate these input-output pairs. Characteristically such algorithm should work in real time and improve the model upon newly available measurements. This well developed broad field includes geometric approaches to stochastic systems [see ► [Stochastic Noises, Observation, Identification and Realization with](#)] as well as more formal learning theories [see ► [Learning, System Identification, and Complexity](#)].

Systems and control theory is unified by a common core of questions, but employs a diverse array of models. Consequently it interfaces a wide range of mathematical disciplines, utilizes many different mathematical tools, but also fosters new development in diverse sub-disciplines. An area that has evolved closely together with control is nonsmooth analysis with its rich theory of tangent objects to nonsmooth sets. Not only are controlled dynamical systems naturally modeled by differential inclusions, but control intrinsically is full of objects that are not differentiable in a classical sense. Most important is the value function in optimal control which encodes the minimal cost (time) to the target. Closely related to classical dynamic programming, this function ought to satisfy the Hamilton–Jacobi–Bellmann equation, a first order partial differential equation, but the value function generally is not differentiable in a traditional sense [see ► [Nonsmooth Analysis in Systems and Control Theory](#)]. Complementary to the nonsmooth approach are differential ge-

ometric approaches which include Lie theory and symplectic geometry. Closely related is a functional analytic operator calculus introduced into control by Agrachëv and Gamkrelidze which linearizes problems by imbedding them into infinite dimensional settings. This dramatically facilitates formal manipulations of nonlinear objects, and also gives new ways of looking at the underlying geometry [see ► [Chronological Calculus in Systems and Control Theory](#)].

Many of the early successful engineering implementations of systems and control theory used models with both continuous time and states (differential equations). But with the advent of ubiquitous digital controllers as well as ever broadening fields of applications the theories for increasingly diverse models are becoming well developed, too. These include, systems with or mix of continuous and discrete states (suitable, for example, to model systems with hysteresis) [see ► [Hybrid Control Systems](#)] and discrete time Hamiltonian and Lagrangian systems [see ► [Discrete Control Systems](#)]. There are many new applications with quickly developing theories such as control of biological and biomedical applications, social dynamics, manufacturing systems, financial markets, and many more. Some of the most intense research in recent years has focused on systems with distributed intelligent agents (controllers) each of which has access to only local information [see ► [Robotic Networks, Distributed Algorithms for](#)]. Compared to these emerging areas the geometric theory of mechanical systems is one of the most mature and thriving research areas with scores of exciting new problems. It serves as a role model for the depth of geometric insight and for its well-understood close interconnections with many other disciplines [see ► [Mechanical Systems: Symmetries and Reduction](#)].

## Acknowledgment

This work was partially supported by the National Science Foundation through the grant DMS 05-09030.

---

## Systems Genetics and Complex Traits

GREGORY W. CARTER, AIMÉE M. DUDLEY  
Institute for Systems Biology, Seattle, USA

## Article Outline

Glossary

Definition of the Subject

Introduction

Measuring Genetic Interactions Genome-Wide



## Constructing Genetic Interaction Networks

### Interpreting Genetic Interactions

### Future Directions

### Acknowledgments

### Bibliography

## Glossary

**Allele** A copy or alternate version of a gene.

**Complex phenotype** A trait (or phenotype) caused by polymorphisms in multiple genes.

**Diploid** An organism with two copies of each chromosome.

**Epistasis** A genetic interaction in which the double-mutant phenotype is identical to one of the single-mutants (masking epistasis); sometimes used as a general term for genetic interaction.

**Genetic interaction** A relationship that characterizes how two (or more) genetic perturbations or alleles combine to affect a phenotype.

**Genetic network** A representation of functional relationships between genes, gene products, or gene perturbations, often displayed as a graph.

**Genetic perturbation** The modification of a gene's presence, structure, or activity by a change in the DNA sequence in or near a gene.

**Genetic transformation** Genetic modification of a cell by the introduction of exogenous genetic material.

**Haploid** An organism with one copy of each chromosome.

**Haplotype** A set of polymorphisms that are linked and thus inherited as a unit.

**Homolog** One of two or more genes with similar DNA sequence due to shared ancestry.

**Hybrid** A diploid offspring from the mating of individuals from two distinct populations.

**Intercross** The mating of individuals from two distinct populations.

**Isogenic** A term describing strains of model organisms that are derived from the same individual or inbred line, e.g. the yeast deletion collection, in which each strain differs only by the presence of a specifically engineered mutation (e.g. a gene knock-out).

**Masking epistasis** A genetic interaction in which the double-mutant phenotype is identical to one of the single-mutants.

**Mendelian genetics** The study of traits caused by variations of a single gene.

**Microarray hybridization** Sequence-specific binding of nucleic acid (DNA or RNA) to short nucleic acid sequences tethered to a microarray chip.

**Molecular barcode** A short, unique sequence inserted into DNA that can be used to identify or quantify a gene, gene product, or strain.

**Network motif** A small subnetwork repeated in a larger network, suggesting repeated organization for a specific function.

**Phenotype** An observable property of an organism; sometimes implies an aberrant property, e.g. the mutant has a phenotype.

**Pleiotropy** The phenomenon in which a mutation in a single gene causes multiple phenotypes.

**Quantitative trait locus (QTL)** A chromosomal region linked to a measurable phenotype.

**RNA interference (RNAi)** A method of silencing a gene's expression by introducing a small RNA molecule with sequences complementary to a portion of the gene's mRNA transcript.

**Reciprocal-hemizygote** A hybrid in which one copy of a gene or chromosomal region has been removed.

**Sporulation** The formation of spores in fungi, produced by meiosis.

**Synthetic sickness or lethality (SSL)** A genetic interaction in which two genetic perturbations without individual fitness effects combine to cause a fitness defect or inviability.

**3' Untranslated region (UTR)** The RNA sequence that follows the protein-coding region of a messenger RNA.

## Definition of the Subject

Among the most fundamental problems in biology is deciphering the relationship between genotype and phenotype. Genetics, the study of how the DNA sequence of an individual (genotype) affects an observable characteristic (phenotype), has many properties of a systems science. Organisms contain large number of protein-coding genes, with ~ 500 predicted in the simple bacterium *Mycoplasma genitalium*, ~ 25,000 in the human genome, and over 50,000 in some plants. The relationship between genotype and phenotype can be complex, with multiple traits per gene (pleiotropy) and multiple genes per trait (complex phenotypes). Genes are the blueprints for both RNA transcripts and proteins that interact physically and functionally to produce the emergent behaviors that characterize life. Furthermore, biological systems can be systematically studied by perturbing elements of the system, such as genes or environmental factors, and quantitatively measuring the effects. This research strategy is generally considered a central component of systems bi-

ology [39,40], and here we refer to its application to the genotype-phenotype problem as *systems genetics*.

Understanding a complex system broadly requires one to (1) identify the elements, (2) determine the function of each element, (3) identify and characterize the interactions between elements, and (4) assemble all of this information into a mathematical model that accurately simulates the system and predicts its responses to novel perturbations. In the context of biology, this process usually begins with sequencing an organism's genome and identifying the functional elements, e.g. the genes. There are numerous methods for determining the function(s) of a given gene product, where "function" can describe both the specific activity carried out by the gene product and the role that activity plays in the organism's response to an environmental or developmental stimulus. For example, molecular or biochemical methods may reveal a protein to be enzyme that converts a substrate to a product, while genetic analysis may reveal that this activity is required for growth in the presence of a specific drug. Similarly, there are both molecular and genetic means of determining which gene products interact with each other, by detecting physical interactions (e.g. protein binding) and genetic interactions (e.g. a mutation in one gene modifies the effect of a mutation in a second gene). These interactions vary with respect to the nature, strength, and directionality of the interaction. Integrating these data into a coherent computational model of the system is a major challenge of systems biology, with early examples including a comprehensive analysis of the galactose utilization pathway in yeast [40] and bacterial chemotaxis [70].

Systems genetics approaches to this problem use large-scale, high-throughput methods to decipher the network of gene functions and genetic interactions in an organism. Apart from their genome-wide scale, the data and methods used often mimic those of conventional genetic analysis. In classical Mendelian genetics, genes are defined as the entities by which traits are inherited. A specific trait exhibited by an organism thus depends on the version(s) of the corresponding gene, known as an allele, carried by that individual. Modern molecular biology has narrowed the concept of a gene to a distinct DNA sequence that is transcribed to encode an RNA or protein product. However, few phenotypes are entirely determined by a single gene. In most cases, multiple genes interact to confer a phenotype. These interactions can vary from simple modifier relationships, in which an allele of one gene alters the phenotype of another gene, to multiple genes with complicated interdependencies. The study of how multiple genes interact to influence a trait is generally known as complex genetics.

Understanding complex genetics is of increasing relevance to the study of human health and is essential to the development of predictive, preventive, and personalized medicine. Many diseases, such as asthma, diabetes, heart disease, and cancer, show a degree of heritability that cannot be traced to a single gene [3]. Moreover, risks inherited from identified disease genes are often modified by background gene variation [11,22,33,34]. Genetic risk in these cases is determined by an individual's allelic profile across many genes, combined with multiple environmental factors. Many of these allelic variations consistently appear together, as haplotypes, which are being mapped in the human HapMap project [1]. Multiple genes involved in cancer are being cataloged in The Cancer Genome Atlas [17]. This data must be combined with new modeling techniques to develop models for understanding the multigenic, molecular basis of human diseases. Developing new treatments for diseases with genetic susceptibilities will require not only the ability to genotype and classify patients on the basis of molecular fingerprints in tissues, but also an understanding of how genetic variants interact to affect clinical outcomes.

## Introduction

Genetic interaction refers to the phenotypic effect of combining two or more genes with allelic variations. Genetic interactions are ideally observed under constant environmental conditions and in isogenic strains of model organisms, so that the lone variables are the genetic perturbations (i.e. the allele forms) carried by the strains being compared. Since genetic variants often combine in complex ways to affect phenotypes, it is necessary to infer and model the functional interactions between trait genes rather than viewing each gene as an independent factor. These functional interactions include all modes of activation or repression of one gene's activity by another. Take as an example, a measurable phenotype and a wild-type strain that defines the genetic background. Next consider two genes, A and B, which can be varied from that of the wild-type strain by mutations such that the (isogenic) genetic background is otherwise unchanged. Taking cell growth in the presence of the drug caffeine as an example phenotype, imagine that a mutation in A inhibits growth in the presence of caffeine. The presence of an additional mutation (B) in the strain (i.e. an AB double mutant) could have one of many different effects. B could suppress the phenotype of A (restoring wild-type growth on caffeine), enhance the phenotype (decreasing growth on caffeine further than that observed in the A single mutant), or show no effect at all (growth equal to that of the A

mutant). Interactions such as these have historically been used to map functional pathways [36].

In addition to the direction of the effect, the magnitude of phenotypic suppression or enhancement also provides important information about the nature of the genetic interaction. Additive effects, where the magnitude of an AB double mutant is equal to the combined effect of each individual's phenotype, are consistent with independent (non-interacting) genes in the same genetic pathway (e. g. where A and B both confer a phenotype) or in different genetic pathways (e. g. where A displays the phenotype, but B does not). Synergistic effects describe an AB double mutant trait that is quantitatively more extreme than would be expected from the linear combination of both. Perhaps the best studied class of synergistic interaction is synthetic lethality [21], in which the combination of two non-lethal mutations, A and B, combine to confer lethality in the AB double mutant. The concept has been appreciated for many years in both fruit fly [75] and yeast [30], and synthetic lethal interactions have identified redundant functions in a common (essential) biological pathways and parallel functions in separate pathways that are able to buffer each other [32]. Such buffering against genetic defects has long been believed to confer a fitness advantage [84]. The buffering of disease-related genes is of particular interest in identifying alleles that either alleviate or aggravate effects within individuals across a genetically diverse population.

In addition to assigning genes to functional pathways and uncovering functional relationships between pathways, the direction and magnitude of genetic interactions can also be used to determine the fine-structure of a genetic network and information flow through the system. Analysis of synthetic and epistatic interactions (Fig. 1) can be used to determine functional topology and the direction of information flow, commonly referred to by geneticists as “ordering genes in a pathway” [4]. Taking again the example of a mutation in gene A that inhibits growth in the presence of caffeine, now consider a mutation in another gene, C, that enhances growth under that condition. If the phenotype of an AC double mutant is a caffeine growth defect equal to that of the A single mutant, A is said to be epistatic to C. Such an interaction is often interpreted as evidence that A acts downstream of C, in that its mutation masks any modifications that occur further upstream of the phenotypic output [4]. This hypothesis is an example of serial information flow through the system, with the perturbed genes influencing each other and the growth phenotype in a linear sequence.

While genetic interaction analysis cannot, when taken alone, decipher the biochemical activity of the gene prod-

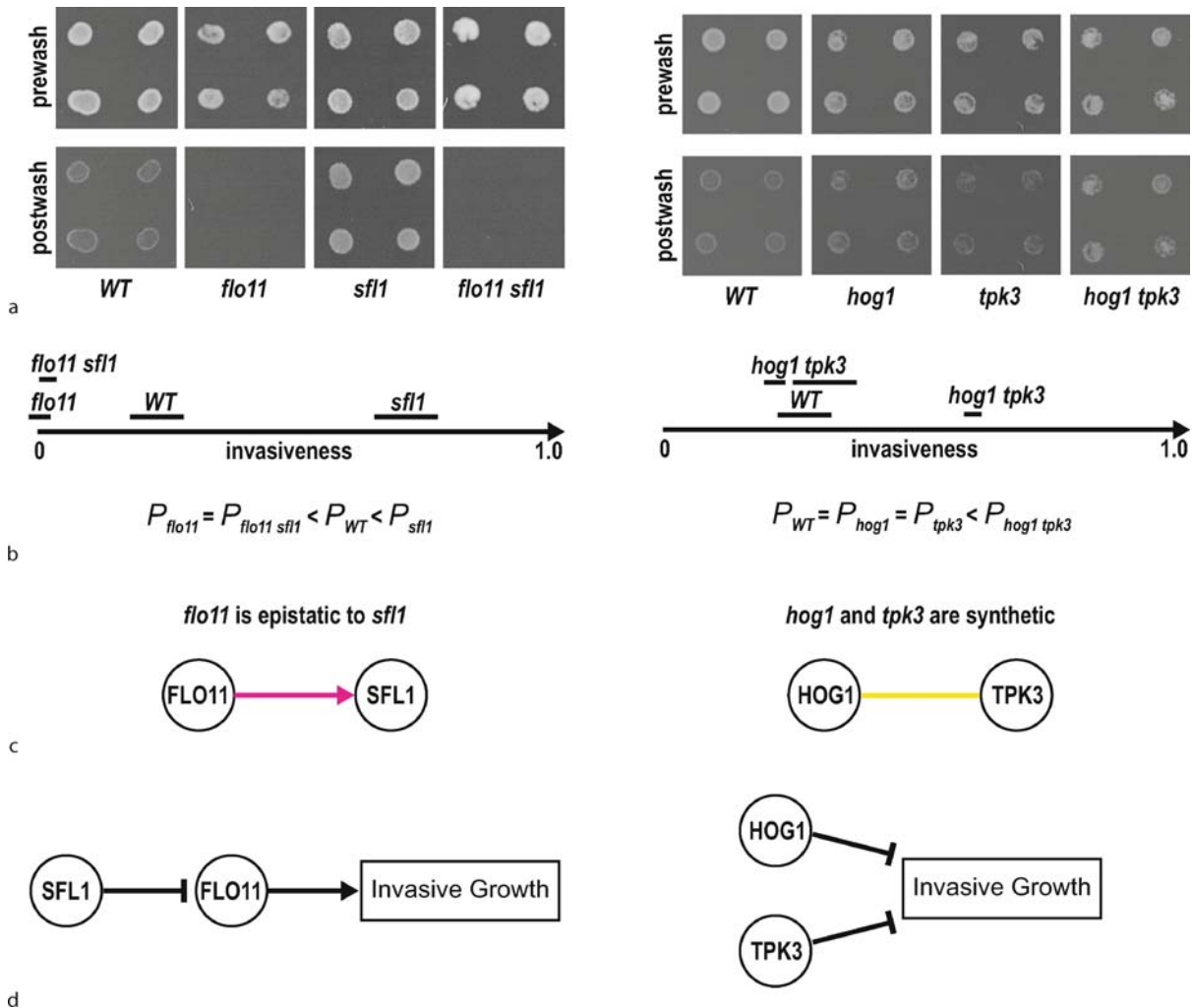
ucts, it can establish functional relationships between gene pairs. One can thus view a genetic interaction as providing a comprehensive view of the coordinated activity of two genes, encompassing every step in cellular processing between genotype and phenotype. A broad picture of complex genetic control emerges when these interactions are assembled into a network.

### Measuring Genetic Interactions Genome-Wide

Since assaying a genetic interaction requires measuring the phenotype of interest in multiple strains, inferring genetic interactions on a genome-wide scale has historically been intractable. The application of high-throughput assays in model organisms with sequenced genomes has made the systematic study of complex genetics possible. To date, these studies have primarily involved the yeast *Saccharomyces cerevisiae*, which is easy to genetically modify and has a large community of researchers committed to the public dissemination of genome-wide reagent sets, and the nematode worm *Caenorhabditis elegans*, which has proven amenable to perturbations by RNAi. The large-scale study of how a genotype contributes to the control of a phenotype has been greatly aided by advances in methods to both systematically perturb genes and measure phenotypes on a genomic scale. These high-throughput methods produce large, quantitative data sets, requiring parallel development of computational and numerical modeling methods to interpret the output in terms of biological function. The primary objective of these methods is to identify functionally relevant genes and describe how they influence one another to generate cellular activity. Using a systems biology approach to integrate this knowledge with other data types has the potential to generate models capable of validating biological insights and identifying high-priority candidate molecules for targeted therapeutic intervention.

### Synthetic Sickness and Lethality

Functional profiling of individual genes is most directly carried out by deleting the gene and observing phenotypic consequences. Through advances in genomics [87], a library of yeast mutants [27] now exists with deletions for every gene in the background of the laboratory strain derived from S288c [86], which was sequenced in the yeast genome project [29]. The yeast genome contains ~6000 protein coding genes (also known as open reading frames or ORFs). Approximately 5000 of these genes are nonessential and were deleted by replacing the gene of interest with a “marker gene” that confers drug resistance and two molecular barcodes, short DNA sequences that uniquely identify each deletion mutant. The remain-



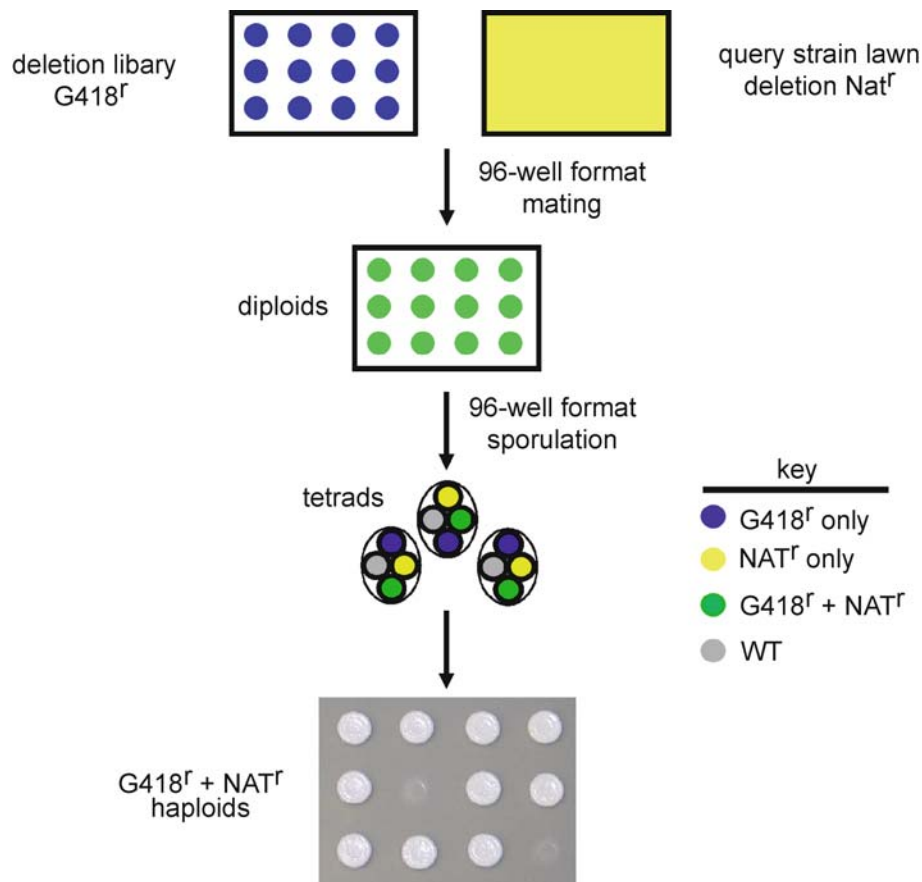
#### Systems Genetics and Complex Traits, Figure 1

Two examples of genetic interactions for yeast invasion. *Left* panels show the interaction between deletions of *SFL1* and *FLO11*, and the *right* panels show the interaction between deletions of *HOG1* and *TPK3*. **a** Yeast invasion assays for relevant strains. Yeast colonies (light spots) were grown for two days on nitrogen-limiting media (prewash) and cells were scraped from the surface (postwash). **b** Quantification of invasive growth, as the ratio of photo-assay signal before and after wash. Experimental ranges were determined by the variation among replicate colonies. **c** Genetic interactions as classified by Drees et al. [23]. **d** Classical genetics interpretations of pathway ordering. Epistasis (*left*) implies the genes are sequential in a pathway, and synthesis (*right*) implies the genes lie in parallel pathways. (Adapted from [23])

ing ~ 1000 genes are essential for cell viability under standard laboratory conditions and thus must be maintained in strains that contain another functional copy of the gene. The lists of essential and non-essential genes are continually updated [44] as ORF annotation and verification improves [46] and closely-related strains of *S. cerevisiae* that have slight variations in essential genes are discovered [63]. Nonetheless, the observation that only 15–20% of yeast genes are essential, at least under optimal growth conditions, suggests that the genetic network governing

yeast growth buffers the cell against genetic variation [32] and implies that multiple genetic mutations may compromise cell growth to a much greater degree than that observed for any single mutation.

Using the yeast deletion library as a starting point, three methods – synthetic genetic array (SGA), diploid-based synthetic lethality analysis with microarrays (dSLAM), and epistatic miniarray profiles (E-MAP) – have been developed as comprehensive, high-throughput assays for detecting pair-wise genetic interactions that



#### Systems Genetics and Complex Traits, Figure 2

Overview of the synthetic genetic array (SGA) and diploid-based synthetic lethality with microarrays (dSLAM) methods. Many of the experimental details have been simplified, more detailed descriptions are presented in [6]. In the SGA method a haploid strain (yellow) is constructed that contains a deletion of the “query gene” marked by a selectable drug marker (Nat<sup>r</sup>) and a specialized reporter gene that facilitates high-throughput selection of haploid progeny that have mated and undergone sporulation [79]. The query strain is mated to an ordered grid, haploid yeast deletion library (blue) marked by the selectable drug marker G418<sup>r</sup>, forming diploids (green) that contain both deletion mutants, but which also contain a functional copy of each gene. Those diploids are then sporulated, allowing random shuffling of the genes and produces haploid progeny, which will only contain one copy of each gene of interest, i. e. the deleted or functional copy of the query gene. In the final step (photograph of actual yeast grids), haploid progeny that contain both the query gene (Nat<sup>r</sup>) and the deletion strain present in that position of the grid (G418<sup>r</sup>) are selected. The ability of the haploid progeny to grow in the presence of both drugs indicates that both deletions are present in the same strain, i. e. no interaction. Lack of growth indicates the inability to recover the double mutant combination, i. e. synthetic lethality

confer synthetic growth defects. Experimental details of these methods have been extensively reviewed [6,48,56]. We briefly summarize them here, as data generated by these methods dominates the genetic interaction data currently available. We then discuss other methods now being applied to comprehensively measure genetic interactions in yeast, worm, and mouse.

The synthetic genetic array (SGA) method (Fig. 2) was developed by Boone and colleagues [79,80] and extended to incorporate essential genes by transcriptional control via an artificial promoter [20]. The availability of a vari-

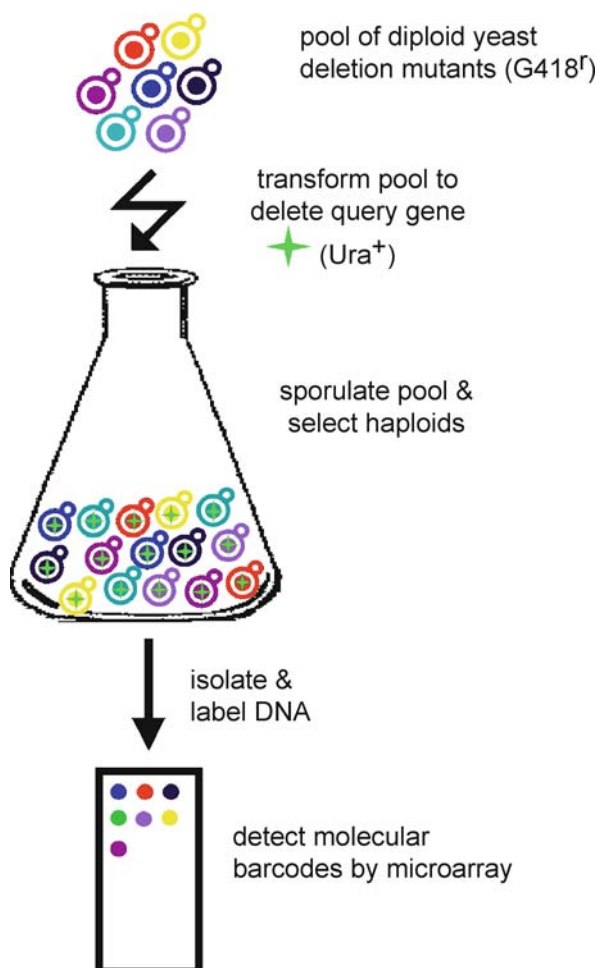
ety of drug-resistance markers enables a simple strategy for efficient construction of double gene deletion strains in yeast. Each strain in the haploid yeast deletion library harbors a gene (Kan) that confers resistance to the drug G418. In the SGA method a second haploid strain is constructed. This strain is a different mating type, and contains a deletion of the gene of interest marked by a gene (Nat) encoding resistance to a second drug (nourseothricin) and a reporter gene that enables selection for haploid progeny strains. This strain is mated to the deletion library, the resulting diploid strains are sporulated, and haploid progeny



resistant to both drugs are selected. Synthetic lethality is indicated by the inability to recover haploid strains with both drug markers, suggesting that this combination of mutant alleles is inviable. Double mutants which show significant growth defects (so called “synthetic sickness”) can also be identified by this method. The throughput of the method is increased by robotic manipulation of strains and phenotype scoring by image analysis software. To date, over 400,000 interactions have been tested and tens of thousands have been discovered [61], providing sufficient data for sophisticated network analysis and a valuable source of functional relationships between hundreds of yeast genes [80].

The issue of precise growth quantification of double-mutants has been addressed by a modification of the SGA method. In this approach, called epistatic miniarray profiles (E-MAP), growth rates are quantified and genetic interactions are identified as cases in which a double mutant strain exhibits a growth rate significantly above or below the expectation based on the two single mutants [19,64]. Essential genes were tested in this study by decreasing mRNA abundance via destabilization of the 3' UTR. Precise quantification of growth rates was performed by photo-assays of colony size and, based on the assumption that genetic interactions are rare, identifying the most significant deviations from the expectation that a double-mutant growth rate will be the product of the two single-mutant growth rates. Interactions with expected growth rates were classified as neutral, those with growth rates above expectation were labeled alleviating, and those with growth rates below expectation were labeled aggravating. To date, pair-wise interactions between 424 genes that localized to the endoplasmic reticulum [64] and 743 *Saccharomyces cerevisiae* genes involved in various aspects of chromosome biology (e.g. DNA replication/repair, chromosome segregation and transcriptional regulation) [18] have been tested.

Diploid-based synthetic lethality analysis with microarrays (dSLAM) (Fig. 3) is an alternate, transformation-based approach to assaying synthetic lethal interactions [57,58]. The method relies on molecular barcodes that uniquely mark each deletion strain to identify double-mutant fitness defects by DNA microarray hybridization. Heterozygous diploid knockouts [27] are initially transformed with a reporter gene that enables their conversion to haploid mutants following sporulation. Then, the gene of interest is deleted via high-efficiency integrative transformation, and the pool of yeast strains is sporulated and separated into single and double mutants using the drug resistance markers. Genomic DNA containing identifying bar codes from each pool is isolated. Bar codes are am-



Systems Genetics and Complex Traits, Figure 3

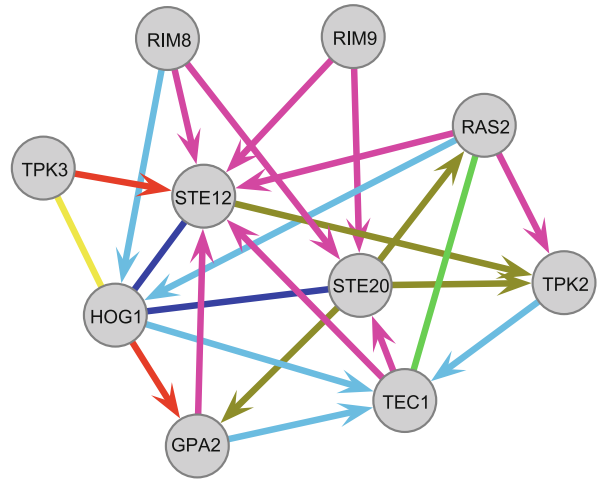
The dSLAM method takes advantage of the molecular barcodes present in each deletion strain to facilitate the manipulation of strains in a pooled format. This method uses a diploid version of the yeast deletion set (heterozygous diploids) in which each strain contains both the gene deletion ( $G418^r$ ) and a functional copy of the same gene. Prior to the analysis, the same reporter gene that facilitates high-throughput selection of haploid progeny used in SGA is introduced into all of the strains (not shown). The first step is the deletion of the query gene in (ideally) all the strains in the pool by transforming in a construct that will delete the gene of interest and contains a marker that will allow successfully transformed cells to grow in the absence of the nutrient uracil ( $Ura^+$ ). The pool is then sporulated and haploid progeny containing both the  $G418^r$  deletion from the collection and  $Ura^+$  query deletion are selected. DNA is then isolated from the pool of strains and the sequences containing the molecular barcodes are fluorescently labeled and hybridized to a DNA microarray that can detect the barcode sequences. Thus, microarray intensity provides a quantitative readout of the amount of each mutant recovered in the selection, with extremely low microarray intensity indicating synthetic lethality

plified, labeled, and hybridized to microarrays to quantify the ratio of single to double mutant growth rates. SSL interactions, also referred to as synthetic fitness or lethality (SFL) defects, are identified by high ratios that correspond to reduced double-mutant growth. Initial comparisons provide evidence that the dSLAM method achieves higher data consistency and lower false-negative rates than the SGA approach. dSLAM can also be modified to study dosage dependent SSL, genetic suppression, and haplo-insufficiency. Although the microarray based method potentially allows a more quantitative measure of synthetic sickness, the low signal intensity observed for many barcodes has hindered precise quantification. Because it relies on growth in mutant pools, the dSLAM method may be the most efficient method for genome-wide SSL screens.

The relatively recent introduction of RNA interference (RNAi) technology for gene knock-down allows multiple loss-of-function genetic perturbations to be carried out at high throughput [43,50] in model organisms such as the worm *C. elegans* and the fruit fly *D. melanogaster*. Large-scale RNAi based synthetic lethal screens have been performed in *C. elegans* [49]. Approximately 350 synthetic interactions between 160 genes in the worm EGF/Ras, Notch, and Wnt signaling pathways were discovered among 65,000 tested gene pairs. The highly-connected “hub” genes encode chromatin regulation proteins that are conserved across multiple species, which were hypothesized to be general buffers of genetic variation. Another study used the same experimental techniques to test functional redundancy between duplicated genes in *C. elegans* [78]. Synthetic interactions were detected in eleven percent of the duplicated pairs tested (16/143), suggesting that in some cases duplicated genes have a degree of functional redundancy. Interestingly, the majority of the interacting gene pairs were found to be duplicated over 80 million years ago, strongly suggesting that this functional redundancy can be conserved by positive selection.

### Genetic Interactions for Quantitative Phenotypes

The vast majority of genetic interactions detected by the large-scale methods described above have been limited to the phenotype space of growth in rich yeast medium. However, phenotypes can be derived from any observable trait, including growth under a variety of environmental conditions, cell morphology, metabolite production, or gene expression pattern. To date, the most comprehensive analysis of genetic interactions detected using a phenotype other than growth in rich media include an analysis of interactions in nitrogen-poor media [23], or in the presence of a DNA damaging agent [71].



**Systems Genetics and Complex Traits, Figure 4**

Subsection of the yeast invasion network showing genes from various pathways and the variety of genetic interactions observed. Edge colors correspond to different interaction modes, e.g. violet is epistasis, yellow is synthesis, and dark blue is additivity. Subset of data from [23]

A mixture of gene deletions and high-copy, gain-of-function gene mutations were combined to map interactions in the yeast invasion network for nitrogen-poor media (Fig. 4) [23]. When starved for nitrogen, many strains of *S. cerevisiae* undergo cell differentiation into a pathogen-like, filamentous growth form and colonies subsequently invade solid media [28]. Drees and co-workers measured invasive growth by photographing colonies before and after vigorously washing cells from the agar surface (Fig. 1a). Multiple replicates of each genotype were assessed using image analysis software to quantitatively determine a mean invasion score and its variance (Fig. 1b). Since the phenotypes of the original single mutants ranged from non-invasive to strongly hyper-invasive, a wide variety of genetic interactions were possible. These interactions were first translated into mathematical inequalities that defined the relationship between the wild-type strain, the two single mutants, and the double mutant in terms of “less than”, “greater than”, and “equal to” relationships. For example, a synthetic interaction in which a double mutant (AB) exhibits a phenotype not shown by either single mutant (A or B) would be either “ $AB < A = B = WT$ ” or “ $A = B = WT < AB$ ”. After systematically enumerating all 45 possible inequalities, the equations were grouped into nine “modes” of genetic interaction. Many of these interaction modes, such as synthesis, epistasis, or suppression, are well known to geneticists, but less familiar modes also appear. For example, the mode labeled “asynthesis”

was defined as a double mutant phenotype which matched that of both single mutants but was distinct from the wild type. This general analysis was applied to over 130 genes and over 1800 pair-wise interactions were tested, revealing a rich spectrum of functional interactions that included examples of all nine modes.

A similarly multi-modal genetic interaction network was more recently constructed for genes that impart resistance to the DNA-damaging agent methyl methanesulfonate (MMS) [71]. Twenty-six such gene deletions (out of over 4700 tested) were identified by reduced growth rate in the presence of MMS, and double-mutant strains were constructed and assayed for all 325 pair-wise combinations. Although the selection method precluded the detection of synthetic interactions, this study is a quantitative analysis of genetic interactions among a set of genes implicated in a specific biological process.

### Analysis of Complex Phenotypes that Result from Natural Variation

Although the majority of systems biology approaches to complex genetics have involved engineered genetic perturbations, genetic interaction studies are being extended to models of natural populations. A series of papers by Kruglyak and colleagues have attempted to infer genetic interactions in recombinant progeny strains [7,8,74]. In this work, two parent yeast strains, the sequenced “laboratory strain” (BY) and a “wild strain” isolated from a vineyard (RM) were mated and 112 haploid progeny strains were isolated. Each strain was genotyped at approximately 3000 markers and RNA expression levels were measured using DNA microarrays [9]. Statistical linkages were found between expression variations and quantitative trait loci (QTLs), and these loci were then assessed for interactions [74]. Although the data suggest that most transcripts were influenced by multiple loci [7], statistically significant pair-wise interactions have only been found for 225 transcripts [8]. Another series of yeast studies, taking high-temperature growth (HTG) as a quantitative phenotype, used reciprocal-hemizygosity as a tool to dissect QTLs in yeast [72]. The contribution of each allele to HTG was inferred by constructing hybrids of two yeast strains and then systematically deleting one copy of each gene within a candidate QTL. The QTL architecture was found to be more complex than expected, featuring both *in cis* and *in trans* linkages and suggesting that QTL regions cannot always be narrowed to a single local gene. The method was also used to explore how the influences of three trait genes were affected by variation in genetic backgrounds [69]. Reciprocal-hemizygosity analysis with ten

other yeast strains demonstrated that the trait genes identified in one background did not necessarily contribute in the same way in other genetic backgrounds. These results demonstrate that, in addition to complexity within a QTL region, there is additional complexity due to genetic interactions. Together these studies illustrate some of the substantial complications that will arise in the analysis of complex genetics in the context of natural populations, such as the identification of trait loci in a background of genetic random variance, the separation of phenotypic effects from *cis*-acting and *trans*-acting polymorphisms, and the limitations of the pair-wise interaction analysis to complex genetics.

Furthermore, the logic of genetic interactions could be applied to functional interactions between different types of biomolecules. For example, recent work has studied the genetic basis of drug response in yeast [59]. Multiple polymorphisms were found that linked genetic loci to certain small molecule drugs, constituting a set of 124 gene-drug interactions. This approach provides a model for understanding variation in drug response within a genetically diverse population. Another study directly applied genetic interaction techniques to pair-wise interactions between drugs by mapping a set of compounds into classes that correspond to the cellular functions they affect [93]. These works begin to address how the understanding of complex genetics can be extended to involve additional system elements and, ultimately, generate a view of the organism as a complex system.

Overall, the study of genetic interactions is a field that has seen substantial development in the past decade. High-throughput methods will continue to open new avenues of research and expand established approaches to genome-wide scales. Data generated by these experiments will require formal and systematic approaches for analysis. Currently, network methods are the dominant framework.

### Constructing Genetic Interaction Networks

Prior to high-throughput experimentation, genetic studies mapped interactions between a few genes in isolated functional pathways. These pathways represent information flow through a small number of elements that regulate a specific cell behavior, such as response to osmotic stress. Today, high-throughput techniques enable the assembly of large-scale genetic interaction data into networks and hold the potential for constructing semi-global maps of multiple, inter-connected pathways.

This wealth of data requires a formalism to efficiently analyze relationships between system elements. Complex systems are often best visualized as networks in which in-

dividual points, called *nodes*, are connected by lines or arrows, called *edges*. In genetic interaction networks, the nodes commonly represent genes, alleles, or gene perturbations (e.g. a gene deletion). Edges that connect nodes can take on different meanings, ranging from classic regulatory influences of one gene's activation or repression of another's activity to edges that signify a direct observation, such as synthetic lethality. The central challenge is to represent the richness of complex genetics in an informative way. Because the construction of each study is informed by a different biological problem, diverse network approaches have been pursued in the literature.

Since they are binary, synthetic lethal interactions are easily represented as a network of gene deletions connected by SSL edges. Although the formal information content of the edges is low, SSL networks have the advantage of relatively high coverage. Pair-wise combination of about 1000 yeast genes have been tested, and although this represents only 3% of all possible yeast pairs it encompasses 500,000 experiments. Four thousand of these pairs were SSL positive, producing a sparse but topologically interesting network. Non-essential genes averaged 34 synthetic lethal partners, and essential genes were found to have more interaction partners on average [80]. Extrapolating these results genome-wide, the authors estimate 200,000 essential pair-wise combinations in yeast, a number of lethal outcomes that is much larger than the 1000 genes that are essential individually [6]. The current network of 4000 genetic interactions demonstrates that SSL interactions are more likely to occur between non-essential genes of the same or related function, unlike SSL partners of essential genes that tend to show a broad spectrum of functions. This suggests that network location can be a predictor of gene function for non-essential genes, enabling novel functional hypotheses for genes based on a common function of interaction partners.

The known SSL network exhibits two interesting topological properties. First, it appears to be a *small world* network, defined as a network in which the length of the shortest path between two nodes is smaller than random expectation. Nodes are densely grouped into local clusters in which the interaction partners of any gene also tend to interact. Thus, the most likely candidates for SSL interactions for a given gene are the interaction partners of its known SSL partners, greatly increasing the likelihood of finding additional interaction partners in a targeted screen [80]. The second network property observed is that the number of SSL partners for each gene follows a power-law or *scale-free* distribution in both the yeast [80] and worm [49] data. As a result, most genes have relatively few interactions, and a small number of genes have rela-

tively large number of interactions. This latter set, referred to as hubs, can be viewed as more important for fitness since the organism is most sensitive to pair-wise perturbations involving them. If this property is also a feature of human networks, the hub concept may inform the treatment of disease. Taking cancer as an example, hubs of cancer cell networks might be targeted for perturbation to increase the chance of a SSL interaction with a gene already genetically perturbed in the cancerous cell.

Nearly 5000 additional, and mostly novel, synthetic fitness or SSL interactions were discovered by studying genes involved in processes related to DNA integrity in yeast using dSLAM [57]. A total of 16 gene modules were identified in the network by clustering genes with a high congruence in SSL interactions, and each module was assigned a function based on its member genes. The resulting network mapped compensatory pathways among multiple biological processes, including DNA replication, DNA repair, checkpoint signaling, chromatin structure maintenance, and the response to oxidative stress. The modular architecture was consistent with prior models, and involved many genes with human homologs linked to cancer and aging.

The study of SSL interaction networks has the practical advantages of a simple, rapid assay and the direct mapping of each positive result to an edge in the interaction network. This narrow definition of genetic interaction, however, cannot explore other types of genetic interaction that have proved informative, such as suppression and masking epistasis. Encompassing all types of genetic interaction requires a systematic formalism such as the inequality relationships developed by Drees et al. [23]. As described above, these inequalities can be grouped into rules of genetic interaction corresponding to familiar and newly defined interaction modes. While many of these modes, such as synthesis, were symmetric with respect to the interacting genes, four of the modes produce directional edges. For example, instances of suppression were rendered with a colored arrow from the suppressor gene to its target gene. The resulting 1800 genetic interactions between 133 genes compose a genetic interaction network, derived with the *PhenotypeGenetics* plug-in for the *Cytoscape* software platform [67]. The nodes of this network included both gene deletions and plasmid-borne over-expressers, and the edges were mapped using nine colors corresponding to the nine modes of genetic interaction based on classical genetic interaction rules. This produced a richly multichromatic network suitable for informatic analysis (Fig. 4).

The multiplicity of genetic interaction modes facilitates the inference of relationships between individual



genes and specific biological functions. In the Drees study, individual alleles were often found to interact in a particular mode, or *monochromatically*, with partner genes of coherent biological function, leading to hypotheses for regulatory and pathway organization [23]. Large-scale patterns of mutual information were also extracted from the data set, and groups of genes with significant mutual information between them formed network cliques corresponding to physical pathways. The genetic interaction patterns define a map of information flow from specific genetic perturbations to quantifiable phenotype effects.

This network was further analyzed by Taylor and Galitski [76], who searched for repeated subnetworks or *motifs* of three or four nodes, in which repetition suggested functional organization. Statistically enriched motifs were then assessed for functional coherence and many instances of functional monochromaticity were observed involving multiple genes. For example, deletions of multiple protein kinase genes were found to suppress gain-of-function perturbations of both signaling and transcription factor genes. This suggests that the kinases mediate information flow from the signaling proteins to transcriptional factors, and these pathways are broken by kinase deletions.

Multi-modal genetic interaction analysis was also performed on the MMS-sensitivity network of St. Onge et al. [71]. Genetic interactions were initially classified as aggravating, alleviating, or neutral interactions based on the double-mutant growth rates relative to that of the two single mutants. Alleviating interactions were of particular interest since there are many ways genes can mask or mimic another's effect. Thus, these interactions were sub-classified into five rules of interaction based on models of pathway ordering and protein co-function, four of which were assigned directionality. The resulting network model uncovered functional relationships and pathway order. The principle of genetic congruence, that genes with similar functional roles will share interaction patterns with functionally related genes in the system, was used to implicate a new regulator in DNA repair.

Both the yeast invasion and MMS networks classified genetic interactions into multiple modes based on interaction inequalities as described above. However, the authors used slightly different classification schemes. For example, in the invasion network “epistasis” included any interaction in which one single-mutant phenotype masks the phenotype of another. In the MMS network a further distinction was drawn between “masking epistasis”, in which the single-mutant with a stronger effect masks the other, and “suppression”, in which a weakly defective single-mutant masks the strongly defective one. While both networks proved informative, it is unclear which choice of

mode classifications are the most meaningful for a given data set. This issue can be analyzed from the perspective of information theory. A measure of network complexity that places information in the context of the system under study could be used for the unsupervised classification of interaction modes for these two data sets. Network classification schemes that produce greater complexity might also generate more biologically meaningful information, i.e. more “biological statements” linking genes to functions were obtained in the more complex networks. Such complexity-based methods represent a powerful and general approach to genetic interaction analysis, with potential for the study of mammalian systems in which interactions are complex and gene annotation data is sparse.

Fully quantitative studies of genetic interactions have been hindered by the limited amount of numeric phenotype data required for more sophisticated mathematical techniques that might decipher the principles underlying genetic interaction. To overcome this limitation many researchers have adopted gene expression microarrays as a quantitative phenotype [9,41,42,47,62,85]. For each strain and environmental condition, genome-wide transcript levels constitute thousands of quantitative phenotype measurements that provide much greater molecular detail than a whole-cell phenotype, such as growth rate. This approach was combined with classical epistasis analysis by Van Driessche et al. [82]. By clustering mutants based on genome-wide expression profiles and applying the principle of masking epistasis, the authors accurately reconstructed the protein kinase A signaling pathway. In a more recent study, gene expression data was combined with whole-cell phenotype data and mathematical modeling to infer and predict genetic interactions relevant for yeast cell differentiation [12]. This analysis used combinatorial perturbations to infer how a set of regulatory genes quantitatively affect each other's activity as well as the expression of thousands of downstream genes. This quantitative model of genetic interaction in terms of genetic influences allowed the prediction of the effects of novel pair-wise genetic perturbations. The decreasing costs of microarrays and advances in genome-wide data analysis make gene expression data an increasingly attractive option for genetic interaction analysis.

### Interpreting Genetic Interactions

The functional relationships described by genetic interaction networks reveal pathway organization and gene function. For predictive and therapeutic purposes, however, it is useful to translate these functional relationships into biomolecular hypotheses. Such hypotheses can sug-



gest additional genetic perturbations to test and validate models, establishing the applicability of network analysis techniques and enabling an iterative process of model refinement. Moreover, the implication of specific alleles, molecules, and interactions in health-related processes can identify candidate genes and gene products for targeted therapeutic intervention with drug compounds or RNA interference. Thus, the interpretation of genetic interaction networks is necessary to improve the basic understanding of complex genetics as well as facilitate the development of personalized and predictive medicine.

Formulating such hypotheses generally involves both molecular data integration and computational modeling. While genetic interactions define functional relationships and uncover pathways of information flow, molecular interaction databases provide a wiring diagram of possible mechanisms for the transmission of functional information. These complementary data types can be integrated to generate models of phenotype regulation [13,31,83,88]. Various computational methods enable this data integration. Kinetic modeling, for example, uses differential equations to predict the functional outcomes when a subset of physical interactions is active. Constraint based approaches use data analysis techniques to narrow the possible mechanisms that may have generated an observation, until the best hypothesis is identified [5,65,89,91]. Finally, whole-cell phenotypes can be dissected at the molecular level by integrating RNA expression levels and proteomic data [12,35,40]. These systems biology approaches have recently been applied to interpret complex genetic networks.

One particularly model-driven approach involves the use of computational models that simulate cellular processes to predict genetic interactions *in silico*. This requires a model, often highly detailed, that captures the relevant activity of every gene product involved in the process being simulated. Metabolic models based on Flux-Balance Analysis (FBA) have been successful in predicting growth rates for yeast [26] and bacteria [25] under a range of environmental conditions and genetic perturbations. Segré and collaborators used an FBA model to explore genetic interactions among metabolic genes in yeast by performing single and double gene knockouts *in silico* [66]. The model predicted that genetic interactions for growth rates fall into three distinct modes: cases in which the double deletion effect is significantly less than the product of the two single mutants (alleviating); cases in which the double deletion effect is significantly greater than the product of the two single mutants (aggravating); and cases in which the double deletion is not significantly different than the product of the two single mutants (neutral). The study also predicted similarity in interaction modes, or monochro-

maticity, between co-functional genes, a result that echoes the finding that genes with common functions preferentially interact in SSL networks. This work illustrates the power of using a system-wide computational simulation to predict complex genetic interactions using sensitive, quantitative whole-cell phenotypes. One of the major disadvantages of such an approach is the extensive amount of information necessary to formulate an FBA model. Furthermore, these predictions have not been empirically verified, and this would involve a major experimental undertaking.

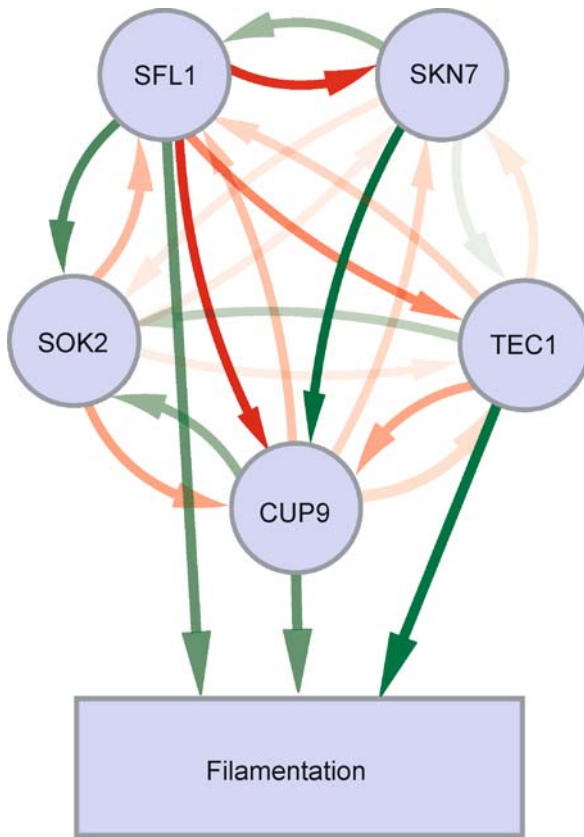
As an alternative to model-based approaches, data-driven computational analysis has also been used to predict interacting gene pairs. Multiple types of relationships between gene products, such as common localization, function, and patterns of gene expression, were used to create a decision tree for predicting SSL interactions in yeast [88]. Genetic interactions were also predicted on a genome-wide scale in the nematode worm *Caenorhabditis elegans* by computationally integrating physical interactions, genetic interaction, gene expression, phenotype, and functional annotation data from yeast, worm, and fly [95]. The authors of this study used logistic regression to determine the predictive power of each type of data, or *feature*, based on a training set of 1816 known genetic interactions in *C. elegans*. Each feature was thereby weighted according to its predictive importance and the probability of interaction between every pair of *C. elegans* genes was calculated. In total, this procedure generated over 18,000 likely genetic interactions. As with previous studies, co-functional genes proved more likely to exhibit genetic interactions (although this is not an entirely independent result given that known genetic interactions between co-functional homologs were the predictors with greatest weight in the logistic regression). A handful of interaction candidates were experimentally tested using RNAi, and about half of the pair-wise perturbations exhibited some degree of genetic interaction in either a vulval development or pharyngeal pumping phenotypes. Interestingly, the vast majority of newly implicated genes did not have vulval development phenotypes when perturbed individually, supporting the hypothesis that the observed double-perturbation phenotypes were the result of genetic interaction. The trade-off inherent in this approach to predicting genetic interactions is that although the analysis can be performed on a genome-wide scale, the nature of each interaction (e.g. additivity, suppression) cannot be predicted. Nevertheless, the approach enables the computational prioritization of potential interaction pairs and greatly reduces the space of likely genetic interactions.

Targeted strategies of data integration have also been used to provide biological context to observed genetic in-

teractions. Kelley and Ideker [45] demonstrated a use for protein-protein and protein-DNA binding networks in understanding the synthetic lethal interaction network in yeast. The study evaluated three standard pathway models for SSL interaction [30,81]: the *between-pathway model*, in which SSL interaction partners are situated in parallel pathways with redundant functions; the *within-pathway model*, in which SSL gene pairs occur for protein subunits of a single complex or pathway that does not lose function until multiple components are removed; and *indirect effects*, in which many pathways are affected and the SSL deletions cannot be associated with specific molecular mechanisms. Based on maximum likelihood, the authors estimated that 40% of yeast SSL interactions could be explained by either the within-pathway or between-pathway models. These models suggest specific patterns of genetic interaction. For the between-pathway model, near-bipartite SSL subnetworks suggest that the missing bipartite links represent undetected interactions. Alternatively, the within-pathway model suggests that genes with many common interaction partners should themselves interact. These predictions were cross-validated by removing one-fifth of the interactions from the data set and then tested to see how well the withheld interactions could be predicted. Both models showed enhanced prediction accuracy relative to randomly predicted interactions, although the between-pathway model was a significantly better predictor, a finding consistent with those suggested by a decision tree approach [88] and an analysis of dSLAM modularity [57,90]. In the worm, integration of protein data also reinforces the compensatory pathway interpretation for SSL interactions [49]. A concurrent study by Zhang et al. [94] integrated molecular and genetic interaction data using the concept of network motifs [53] to guide biological interpretation [92]. In this work, a global network containing protein-protein binding, protein-DNA binding, gene co-expression, sequence homology, and SSL relationships was systematically searched for all possible three-node motifs. A motif in which two proteins linked by protein-protein interaction both have SSL interactions with a third protein was detected in the network ten times more often than random expectation. This was interpreted as a sign of higher-order organization: two proteins that bind in the same complex or pathway share genetic interactions with other proteins in a separate complex or pathway. This interpretation was supported by a search for a four-protein motif in which two parallel pairs of co-binding proteins are linked by all four possible SSL interactions. This motif was found over two hundred times more than random expectation, further supporting the interpretation that SSL interactions are often hallmarks of parallel

or compensatory pathways. The goals of integrative analysis are the systematic interpretation of SSL interactions in terms of molecular interactions and the association of specific pathways with observed genetic interactions.

For networks that include multiple modes of genetic interactions, such as suppression, additivity, and genetic synthesis, finding biologically meaningful, non-random network topologies in an integrated network is hindered by the enormous number of possible motifs and the large datasets required for adequate statistics. One strategy for interpreting these rich genetic interaction networks is a quantitative generalization of the classical genetic interaction approach of observing how genetic perturbations interact to affect phenotypes [4]. This method has historically been used to infer how genes influence one another and thereby affect a downstream phenotype, revealing functional relationships such as activation, repression, and pathway order. Using filamentous growth in yeast as a model system (a precursor to invasive growth due to nitrogen starvation), a data decomposition technique was developed to separate genetically “direct” (not necessarily molecularly direct) effects of regulator genes from the genetically “indirect” effects that involve genetic interactions between regulator genes (Fig. 5) [12]. Molecular interaction data were then integrated with the decomposition results to construct regulatory network models of information flow through paths of genes (and gene products) linked by physical interactions. These networks represented quantitative hypotheses of influence from the causal perturbation, through a series of molecular interactions, to the expression of affected genes. Since additional genes were implicated in specific influence-mediator roles in the network, quantitative predictions could be made for gene expression effects arising from their deletion and combinations of deletions. This was done by removing the quantitative influence corresponding to the molecular pathways involving the deleted gene(s). A second round of gene expression data was collected to test a set of predictions and the results demonstrated genome-wide success. Next, by using additional decomposition methods to link gene expression patterns with the filamentous growth phenotype, a network model of whole-cell phenotype regulation was constructed. Predictions for additional perturbations were formulated and experimentally tested. The model successfully implicated new regulators of the phenotype and was used to accurately predict phenotypes for novel combinatorial deletions. While this initial work used laboratory-engineered mutant strains, the techniques can be readily extended to dissect and map complex genetic systems in which a handful of trait genes have been identified. With the increasing availability of human interaction



Systems Genetics and Complex Traits, Figure 5

Influence network of five transcription factors that are trait genes for the filamentous growth phenotype in yeast. Green and red edges map positive and negative influences, respectively, between genes and from genes to the filamentation phenotype. Edge intensity denotes influence strength. (Adapted from [12])

data [73] and further modeling developments to address allelic variation in outbred populations, similar quantitative and integrative techniques may ultimately be applied to disease-related models.

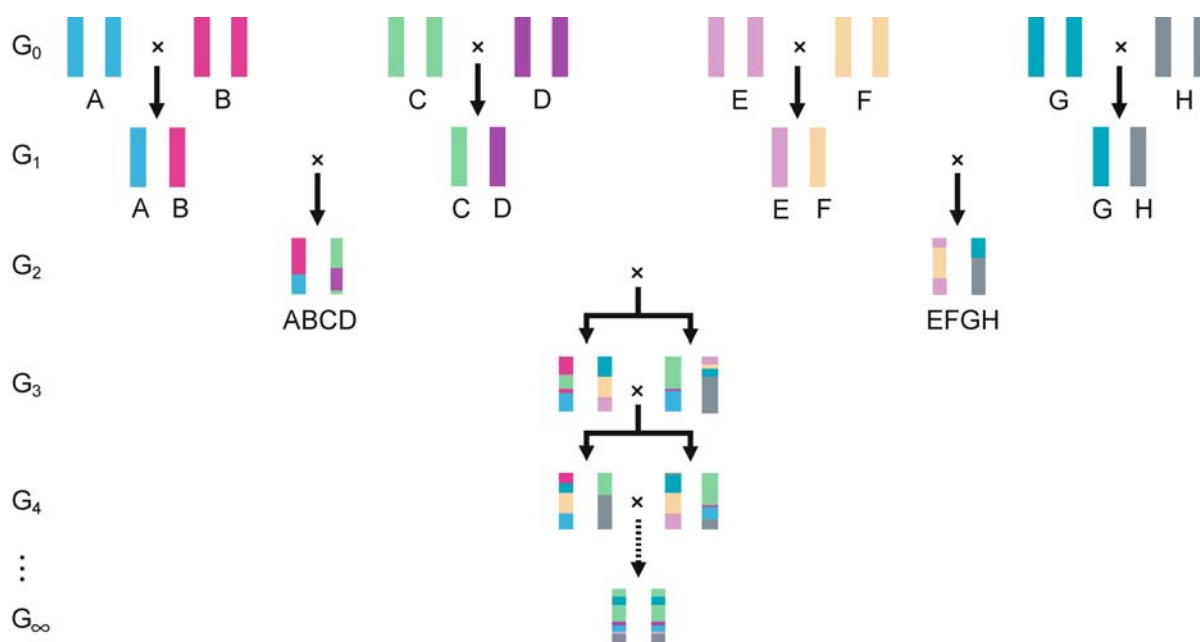
### Future Directions

Although genetic complexity is central to all life, systems biology approaches to complex genetics have to this point focused on simpler model organisms. The genetic manipulability, rapid reproduction, and genome-wide interaction databases for yeast, worm, and model prokaryotes have enabled efficient development of experimental and analytical methods. However, the possibility of understanding human disease through the study of complex genetics has catalyzed substantial advances in genetic association studies in mammalian models and human populations [54,55]. The current status of systems approaches

to human health are discussed elsewhere in this encyclopedia (see ► [Systems Biology, Introduction to](#) and ► [Systems Biology of Human Immunity and Disease](#)), and here we will limit discussion to research opportunities in model systems that are candidates for the systems approaches discussed in this chapter. Future developments can be roughly split into developments in two categories: laboratory resources and computational methods.

Recent programs in rodent breeding have engineered controlled yet genetically diverse populations. These experiments conceptually resemble the analysis of recombinant meiotic progeny generated by crosses in yeast (Sect. “[Measuring Genetic Interactions Genome-Wide](#)”). In mouse, two inbred, homozygous parent strains are crossed and two of the resulting F<sub>1</sub> offspring are mated (Fig. 6). The resulting F<sub>2</sub> progeny are then segregated into breeding pairs, and each pair begins a line of repeated sibling inbreeding that continues until a homozygous inbred strain is obtained. Unlike yeast, in which homozygous diploids can be obtained in a matter of weeks, producing recombinant inbred (RI) lines in mice typically requires 20 generations, or 7–8 years, to produce the desired RI lines. Nevertheless, RI genetics have been integrated with gene expression and phenotype assays in mice [10,15,51] and rats [37]. While each of these studies was able to associate individual loci with gene expression patterns, implicating new genes in metabolic, nervous system, and stem cell function, the relatively small numbers of RI strains (approximately 30 in each study) was insufficient for systematically inferring polygenic effects. Inferring such effects would require additional model parameters to encompass the quantitative interaction between each gene pair, leading to overfitting for small data sets. Nevertheless, these works provide prototypes for further studies using RI strains.

While constructing RI strains is expensive and laborious, they have significant advantages over simpler intercross strains in the study of complex genetics. Multiple mice from each RI line will allow biological replication and the observation of isogenic strains under a variety of environmental conditions. Strains can be assayed over long time scales and multiple generations. A collection of shared RI strains would be especially valuable as a community asset both in terms of labor saved, since the strains need only be constructed and genotyped once, and data sharing, since the results from multiple researchers can be integrated in novel analysis. With this in mind, the Complex Trait Consortium is pursuing the Collaborative Cross project [16], which aims to create approximately 1000 RI strains derived from eight inbred founder strains. This set of RI strains will contain a degree of genetic diversity suf-



**Systems Genetics and Complex Traits, Figure 6**

The collaborative cross strategy for breeding recombinant inbred (RI) strains. The starting parents of the cross (Generation 0,  $G_0$ ) consists of eight inbred laboratory mouse strains. The goal of the breeding strategy is to create 1000 new RI mouse lines that contain different combinations of alleles from the original parents. A theoretical example of the generation of a single chromosome for a single RI strain is shown. In the first generation of the cross ( $G_1$ ) progeny receive an unaltered copy of a chromosome from each parent. In all subsequent generations chromosomal material is exchanged between chromosomes generating chromosomes that contain genomic regions from a variety of parents. The lines are ultimately converted to recombinant inbred (RI) strains, in which an individual contains identical copies of these hybrid chromosomes ( $G_{\text{infinity}}$ ) by multiple generations of inbreeding, in this case mating brother-sister pairs. (Reprinted with permission from The Complex Trait Consortium [16])

ficient to uncover gene-gene, gene-drug, and gene-environment interactions. Furthermore, the resulting 1000 inbred mouse lines could be crossed to produce hundreds of thousands of intercross (RIX) strains. Inbred homozygosity ensures that these genotypes will be reproducible and the genetic differences between RI strains will produce low coefficients of inbreeding. Because the RIX strains are not themselves inbred, they will more closely resemble the human population in terms of heterozygosity, genetic mixing, and hybrid vigor. When combined with the proliferation of gene annotation and molecular interaction data, resources like the mice derived from the Collaborative Cross will allow systems approaches to mammalian genetics and provide a platform for adaption to human disease.

In addition to improvements in laboratory resources, a number of analytical advances are required to better understand large-scale genetic interaction networks. These include further developments in data integration, quantitative prediction, data analysis, ontologies, and frameworks for classifying complex genetics.

Strategies of data integration have taken diverse forms. These range from all-inclusive, nearly universal strategies of data pooling [88,95], to integrating only a handful of physical interactions implicated by phenotype modeling [13,45], to modeling constrained by incomplete data resources [52]. The accuracy and reliability of data varies greatly both across data types (e.g. protein-protein versus SSL interactions) and within data types (e.g. protein-protein data from yeast two-hybrid versus co-immunoprecipitation assays). Some methods have used machine learning techniques to weigh data types based on predictive utility [38,94]. Through the repeated application of such methods to multiple data sets and biological systems a consistent picture of data reliability might emerge, providing valuable information for modeling and interpretation of complex genetic systems. Finally, in many cases the integration of published data appears to be performed at a late stage in the work as researchers seek more complete explanations for their results. While this is to some extent an inevitable result of large-scale screens or discovery-driven science, it would be advantageous for investigators to con-

sider from the outset the types of data that may be most complementary and form a coherent plan of data integration [24]. The use of successful mathematical models and biological simulations at such an early stage of research would aid in this integrative research design.

Genetic interaction analysis is often showcased for its predictive modeling potential. Since genetic interactions inherently reveal functional information, undercharacterized genes can be assigned to biological pathways [23,71] or functions [80], and these implicit hypotheses about gene function can be experimentally tested. However, physical interaction networks organized into pathways of information flow can also explicitly predict the consequences of additional perturbation [12,66]. The predictions are often quantitative and represent a potentially powerful tool for predictive medicine when applied to the development of therapeutic interventions. Expanded use of computational modeling has the potential to generate more accurate and sophisticated predictions of how systems respond to genetic or environmental perturbations. Established modeling strategies include kinetic modeling, which has been used to understand complex cell circuitry [14,60] and to elucidate subtle stochastic effects [2], and Boolean network approaches, which have been used to model sophisticated systems such as host-parasite interactions [77] and uncover fundamental properties of biological dynamics [68]. Combined with time-series assays of responses to environmental stimuli in genetically diverse organisms, these quantitative modeling techniques can be a powerful tool in understanding the dynamics of polygenic traits.

In the analysis of genetic interaction data, the field of systems genetics now is primarily focused on synthetic lethal interactions, to the extent that many authors use the terms genetic interaction and synthetic lethality synonymously. However, synthetic lethal interactions may be less relevant in the complex genetics of natural populations than in laboratory-based screens. Their broad application may also be limited by reliance on technology specific to model organisms and networks composed of binary interactions. For instance, although synthetic screens can reveal compensatory modules and classify genes into functional pathways, the scale-free architecture of these networks often features many pathways connected to a common hub [49]. Since newly discovered synthetic interactions are most likely to include a hub and a non-hub gene, many pathways will be connected to the hub, and thus be equally likely candidates. It may therefore be difficult to place the latter gene in the correct pathway without additional information. Alternatively, considering a broader spectrum of genetic interactions can both resolve pathway

membership, via mutually informative partners, and provide information about pathway structure, e. g. the placement of a factor in a signaling or transcriptional cascade. A network encoding a more general set of genetic interaction rules can thus enable the modeling of specific mechanisms in detail rather than congruence methods based on guilt-by-association. Examples of such rules are directional, targeted interactions like epistasis and suppression (Fig. 4) [23,71] or quantifiable influences describing one gene's influence on another (Fig. 5) [12].

To this point, the classification and interpretation of genetic interactions has been inconsistent due to the paucity of quantitative phenotypes that can be widely screened paired with a limited knowledge of the nature and prevalence of genetic interactions. As the field of complex genetics evolves, so too will the meaning of genetic interaction. To date, genetic interaction studies have primarily focused on pair-wise interactions that are interpreted on a network level. This is often a result of exploring complex genetics with engineered strains, in which single and double genetic perturbations are systematically performed. The combinatorial expansion makes systematic construction of higher-order perturbations (such as triple-mutants) intractable. Indeed, a comprehensive set of double perturbations is daunting, and would be made even more complicated when one considers the corresponding increase in the number and richness of possible phenotypes. In contrast, populations derived from genetically diverse strains, such as yeast and mouse crosses, will sample all genetic combinations of arbitrarily high order. It is possible that network analysis based on pair-wise interactions will be inadequate for experiments involving these strains. Modules or densely interacting “gene complexes” may better describe the space of polygenic effects. With these issues being compounded by the limited capacity to assay many of the phenotypes most relevant to disease, it is currently unknown which approaches will yield the most information about the underlying biology. It is likely that the optimal approach will vary by organism and experiment.

Mature network analyses developed in a broad and systematic framework could be readily expanded beyond genotype-phenotype modeling. To model effects of environment and drugs, additional factors that affect phenotype might be integrated. Potential factors include metabolites, exogenous molecules such as drugs or environmental compounds, and external conditions such as temperature or cell density. A model encompassing multiple factors could quantify how each factor interacts or influences each other factor, in order to produce a map of how effectors interact to generate a phenotype or outcome. Predictive network models focused on specific phe-



notypes would allow the development of highly specific drugs and aid in the prediction of side effects by observing how a particular perturbation affects connected systems *in silico*. These models could also aid in the identification of allelic combinations that cause genetic disease, the prediction of desired drug-allele interactions for personalized medicine, and the development of multi-drug treatments. Creating such models will require substantial advances in integration of mathematical modeling and genetics, with the focus modest-sized systems rather than genome-scale screens.

In addition to advancing biomedicine, predictive models could also inform evolutionary biology and bioengineering. Rare mutations that confer fitness advantages could be identified by systematically predicting the fitness consequences of perturbations. Perturbations predicted to have effects on multiple phenotypes may uncover evolutionary trade-offs, in which the fitness balance might vary as the environment changes. Fitness assays, such as the E-MAP method in yeast [64], could be used to experimentally test these predictions in model organisms. These effects could in turn be used as input for population modeling and predict evolutionary trends suitable for laboratory testing with fast-replicating yeast or bacteria. As a more practical application, the ability to predict the behavior of re-engineered and synthetic biological networks is a key ingredient in the design and development of novel bioengineered systems. Models that predict complex phenotypes quantitatively and precisely will be a necessary component of elaborate modeling schemes that link genotype, phenotype, and ecosystem.

High throughput data acquisition and the development of computational approaches have enabled the field of systems genetics to progress rapidly. Far-sighted experimental planning and continued technological advances can be expected to fuel continued progress. Current studies, primarily focused on yeast as a model system, are laying the groundwork for network analysis and modeling methods that may be adapted to higher organisms. A target for early adaption may lie in the field of metazoan development to study complex genetic networks involved in multicellular biology, possibly expanding the use of *C. elegans* as a model due to its ease of manipulation by RNAi technology [43,48]. Knowledge accrued from these studies coupled with advances in statistical genetics may form a solid basis for applications in more genetically diverse populations, including the RI mouse strains generated by the Collaborative Cross and their RIX intercross counterparts. The result will be new tools for complex human genetics. Meanwhile, the mapping of the human HapMap, The Cancer Genome Atlas, and human interaction net-

works will begin to provide data necessary to make use of these tools in personalized and predictive medicine.

## Acknowledgments

G.W.C. is supported by grants K25GM079404 and P50GM076547 from NIH and FIBR EF-0527023 from NSF. A.M.D. is supported by a Genome Scholar and Faculty Transition award from the NIH/NHGRI.

## Bibliography

1. International HapMap Consortium (2005) A haplotype map of the human genome. *Nature* 437:1299–320
2. Acar M, Becskei A, van Oudenaarden A (2005) Enhancement of cellular memory by reducing stochastic transitions. *Nature* 435:228–32
3. Altmüller J, Palmer LJ, Fischer G, Scherb H, Wjst M (2001) Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet* 69:936–50
4. Avery L, Wasserman S (1992) Ordering gene function: the interpretation of epistasis in regulatory hierarchies. *Trends Genet* 8:312–6
5. Bar-Joseph Z, Gerber GK, Lee TI, Rinaldi NJ, Yoo JY, Robert F, Gordon DB, Fraenkel E, Jaakkola TS, Young RA, Gifford DK (2003) Computational discovery of gene modules and regulatory networks. *Nat Biotechnol* 21:1337–42
6. Boone C, Bussey H, Andrews BJ (2007) Exploring genetic interactions and networks with yeast. *Nat Rev Genet* 8:437–49
7. Brem RB, Kruglyak L (2005) The landscape of genetic complexity across 5700 gene expression traits in yeast. *Proc Natl Acad Sci USA* 102:1572–7
8. Brem RB, Storey JD, Whittle J, Kruglyak L (2005) Genetic interactions between polymorphisms that affect gene expression in yeast. *Nature* 436:701–3
9. Brem RB, Yvert G, Clinton R, Kruglyak L (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* 296:752–5
10. Bystrykh L, Weersing E, Dontje B, Sutton S, Pletcher MT, Wiltshire T, Su AI, Vellenga E, Wang J, Manly KF, Lu L, Chesler EJ, Alberts R, Jansen RC, Williams RW, Cooke MP, de Haan G (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using ‘genetical genomics’. *Nat Genet* 37:225–32
11. Carlborg O, Haley CS (2004) Epistasis: too often neglected in complex trait studies? *Nat Rev Genet* 5:618–25
12. Carter GW, Prinz S, Neou C, Shelby JP, Marzolf B, Thorsson V, Galitski T (2007) Prediction of phenotype and gene expression for combinations of mutations. *Mol Syst Biol* 3:96
13. Carter GW, Rupp S, Fink GR, Galitski T (2006) Disentangling information flow in the Ras-cAMP signaling network. *Genome Res* 16:520–6
14. Chen KC, Calzone L, Csikasz-Nagy A, Cross FR, Novak B, Tyson JJ (2004) Integrative analysis of cell cycle control in budding yeast. *Mol Biol Cell* 15:3841–62
15. Chesler EJ, Lu L, Shou S, Qu Y, Gu J, Wang J, Hsu HC, Mountz JD, Baldwin NE, Langston MA, Threadgill DW, Manly KF, Williams RW (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat Genet* 37:233–42

16. Churchill GA, Airey DC, Allayee H, Angel JM, Attie AD, Beatty J, Beavis WD, Belknap JK, Bennett B, Berrettini W, Bleich A, Bogue M, Broman KW, Buck KJ, Buckler E, Burmeister M, Chesler EJ, Cheverud JM, Clapcote S, Cook MN, Cox RD, Crabbe JC, Crusio WE, Darvasi A, Deschepper CF, Doerge RW, Farber CR, Forejt J, Gaile D, Garlow SJ, Geiger H, Gershenfeld H, Gordon T, Gu J, Gu W, de Haan G, Hayes NL, Heller C, Himmelbauer H, Hitzemann R, Hunter K, Hsu HC, Iraqi FA, Ivandic B, Jacob HJ, Jansen RC, Jepsen KJ, Johnson DK, Johnson TE, Kempermann G, Kendziorski C, Kotb M, Kooy RF, Llamas B, Lammer F, Lassalle JM, Lowenstein PR, Lu L, Lusis A, Manly KF, Marcucio R, Matthews D, Medrano JF, Miller DR, Mittleman G, Mock BA, Mogil JS, Montagutelli X, Morahan G, Morris DG, Mott R, Nadeau JH, Nagase H, Nowakowski RS, O'Hara BF, Osadchuk AV, Page GP, Paigen B, Paigen K, Palmer AA, Pan HJ, Peltonen-Palotie L, Peirce J, Pomp D, Pravenec M, Prows DR, Qi Z, Reeves RH, Roder J, Rosen GD, Schadt EE, Schalkwyk LC, Seltzer Z, Shimomura K, Shou S, Sillanpaa MJ, Siracusa LD, Snoeck HW, Spearow JL, Svenson K, Tarantino LM, Threadgill D, Toth LA, Valdar W, de Villena FP, Warden C, Whatley S, Williams RW, Wiltshire T, Yi N, Zhang D, Zhang M, Zou F (2004) The Collaborative Cross, a community resource for the genetic analysis of complex traits. *Nat Genet* 36:1133–7
17. Collins FS, Barker AD (2007) Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Sci Am* 296:50–7
18. Collins SR, Miller KM, Maas NL, Roguev A, Fillingham J, Chu CS, Schuldiner M, Gebbia M, Recht J, Shales M, Ding H, Xu H, Han J, Ingvarsdottir K, Cheng B, Andrews B, Boone C, Berger SL, Hieter P, Zhang Z, Brown GW, Ingles CJ, Emili A, Allis CD, Toczyski DP, Weissman JS, Greenblatt JF, Krogan NJ (2007) Functional dissection of protein complexes involved in yeast chromosome biology using a genetic interaction map. *Nature* 446:806–10
19. Collins SR, Schuldiner M, Krogan NJ, Weissman JS (2006) A strategy for extracting and analyzing large-scale quantitative epistatic interaction data. *Genome Biol* 7:R63
20. Davierwala AP, Haynes J, Li Z, Brost RL, Robinson MD, Yu L, Mnaimneh S, Ding H, Zhu H, Chen Y, Cheng X, Brown GW, Boone C, Andrews BJ, Hughes TR (2005) The synthetic genetic interaction spectrum of essential genes. *Nat Genet* 37:1147–52
21. Dobzhansky T (1946) Genetics of Natural Populations. Xiii. Recombination and Variability in Populations of *Drosophila Pseudobscura*. *Genetics* 31:269–90
22. Donehower LA, Harvey M, Slagle BL, McArthur MJ, Montgomery CA Jr, Butel JS, Bradley A (1992) Mice deficient for p53 are developmentally normal but susceptible to spontaneous tumours. *Nature* 356:215–21
23. Drees BL, Thorsson V, Carter GW, Rives AW, Raymond MZ, Avila-Campillo I, Shannon P, Galitski T (2005) Derivation of genetic interaction networks from quantitative phenotype data. *Genome Biol* 6:R38
24. Facciotti MT, Bonneau R, Hood L, Baliga NS (2004) Systems Biology Experimental Design – Considerations for Building Predictive Gene Regulatory Network Models for Prokaryotic Systems. *Current Genomics* 5:527–544
25. Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO (2007) A genome-scale metabolic reconstruction for *Escherichia coli* K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 3:121
26. Forster J, Famili I, Fu P, Palsson BO, Nielsen J (2003) Genome-scale reconstruction of the *Saccharomyces cerevisiae* metabolic network. *Genome Res* 13:244–53
27. Giaever G, Chu AM, Ni L, Connelly C, Riles L, Veronneau S, Dow S, Lucau-Danila A, Anderson K, Andre B, Arkin AP, Astromoff A, El-Bakkoury M, Bangham R, Benito R, Brachet S, Campanaro S, Curtiss M, Davis K, Deutschbauer A, Entian KD, Flaherty P, Foury F, Garfinkel DJ, Gerstein M, Gotte D, Guldener U, Hegemann JH, Hempel S, Herman Z, Jaramillo DF, Kelly DE, Kelly SL, Kotter P, LaBonte D, Lamb DC, Lan N, Liang H, Liao H, Liu L, Luo C, Lussier M, Mao R, Menard P, Ooi SL, Revuelta JL, Roberts CJ, Rose M, Ross-Macdonald P, Scherens B, Schimmack G, Shafer B, Shoemaker DD, Sookhai-Mahadeo S, Storms RK, Strathern JN, Valle G, Voet M, Volckaert G, Wang CY, Ward TR, Wilhelm J, Winzeler EA, Yang Y, Yen G, Youngman E, Yu K, Bussey H, Boeke JD, Snyder M, Philippsen P, Davis RW, Johnston M (2002) Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* 418:387–91
28. Gimeno CJ, Ljungdahl PO, Styles CA, Fink GR (1992) Unipolar cell divisions in the yeast *S. cerevisiae* lead to filamentous growth: regulation by starvation and RAS. *Cell* 68:1077–90
29. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M, Louis EJ, Mewes HW, Murakami Y, Philippsen P, Tettelin H, Oliver SG (1996) Life with 6000 genes. *Science* 274:546, 563–7
30. Guarente L (1993) Synthetic enhancement in gene interaction: a genetic tool come of age. *Trends Genet* 9:362–6
31. Gunsalus KC, Ge H, Schetter AJ, Goldberg DS, Han JD, Hao T, Berriz GF, Bertin N, Huang J, Chuang LS, Li N, Mani R, Hyman AA, Sonnichsen B, Echeverri CJ, Roth FP, Vidal M, Piano F (2005) Predictive models of molecular machines involved in *Caenorhabditis elegans* early embryogenesis. *Nature* 436: 861–5
32. Hartman JL, Garvik B, Hartwell L (2001) Principles for the buffering of genetic variation. *Science* 291:1001–1004
33. Harvey M, McArthur MJ, Montgomery CA Jr, Bradley A, Donehower LA (1993) Genetic background alters the spectrum of tumors that develop in p53-deficient mice. *Faseb J* 7:938–43
34. Harvey M, Vogel H, Morris D, Bradley A, Bernstein A, Donehower LA (1995) A mutant p53 transgene accelerates tumour development in heterozygous but not nullizygous p53-deficient mice. *Nat Genet* 9:305–11
35. Haugen AC, Kelley R, Collins JB, Tucker CJ, Deng C, Afshari CA, Brown JM, Ideker T, van Houten B (2004) Integrating phenotypic and expression profiles to map arsenic-response networks. *Genome Biol* 5:R95
36. Huang LS, Sternberg PW (1995) Genetic dissection of developmental pathways. *Methods Cell Biol* 48:97–122
37. Hubner N, Wallace CA, Zimdahl H, Petretto E, Schulz H, Maciver F, Mueller M, Hummel O, Monti J, Zidek V, Musilova A, Kren V, Causton H, Game L, Born G, Schmidt S, Muller A, Cook SA, Kurtz TW, Whittaker J, Pravenec M, Aitman TJ (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat Genet* 37:243–53
38. Hwang D, Rust AG, Ramsey S, Smith JJ, Leslie DM, Weston AD, de Atauri P, Aitchison JD, Hood L, Siegel AF, Bolouri H (2005) A data integration methodology for systems biology. *Proc Natl Acad Sci USA* 102:17296–301
39. Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* 2:343–72

40. Ideker T, Thorsson V, Ranish JA, Christmas R, Buhler J, Eng JK, Bumgarner R, Goodlett DR, Aebersold R, Hood L (2001) Integrated genomic and proteomic analyses of a systematically perturbed metabolic network. *Science* 292:929–34
41. Jansen RC (2003) Studying complex biological systems using multifactorial perturbation. *Nat Rev Genet* 4:145–51
42. Jansen RC, Nap JP (2001) Genetical genomics: the added value from segregation. *Trends Genet* 17:388–91
43. Kamath RS, Fraser AG, Dong Y, Poulin G, Durbin R, Gotta M, Kanapin A, Le Bot N, Moreno S, Sohrmann M, Welchman DP, Zipperlen P, Ahringer J (2003) Systematic functional analysis of the *Caenorhabditis elegans* genome using RNAi. *Nature* 421:231–7
44. Kastenmayer JP, Ni L, Chu A, Kitchen LE, Au WC, Yang H, Carter CD, Wheeler D, Davis RW, Boeke JD, Snyder MA, Basrai MA (2006) Functional genomics of genes with small open reading frames (sORFs) in *S. cerevisiae*. *Genome Res* 16:365–73
45. Kelley R, Ideker T (2005) Systematic interpretation of genetic interactions using protein networks. *Nat Biotechnol* 23:561–6
46. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423:241–54
47. Klose J, Nock C, Herrmann M, Stuhler K, Marcus K, Bluggel M, Krause E, Schalkwyk LC, Rastan S, Brown SD, Bussow K, Himmelbauer H, Lehrach H (2002) Genetic analysis of the mouse brain proteome. *Nat Genet* 30:385–93
48. Lehner B (2007) Modelling genotype-phenotype relationships and human disease with genetic interaction networks. *J Exp Biol* 210:1559–66
49. Lehner B, Crombie C, Tischler J, Fortunato A, Fraser AG (2006) Systematic mapping of genetic interactions in *Caenorhabditis elegans* identifies common modifiers of diverse signaling pathways. *Nat Genet* 38:896–903
50. Lehner B, Tischler J, Fraser AG (2006) RNAi screens in *Caenorhabditis elegans* in a 96-well liquid format and their application to the systematic identification of genetic interactions. *Nat Protoc* 1:1617–20
51. Li H, Chen H, Bao L, Manly KF, Chesler EJ, Lu L, Wang J, Zhou M, Williams RW, Cui Y (2006) Integrative genetic analysis of transcription modules: towards filling the gap between genetic loci and inherited traits. *Hum Mol Genet* 15:481–92
52. Maslov S, Sneppen K, Eriksen KA, Yan KK (2004) Upstream plasticity and downstream robustness in evolution of molecular networks. *BMC Evol Biol* 4:9
53. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U (2002) Network motifs: simple building blocks of complex networks. *Science* 298:824–7
54. Monks SA, Leonardson A, Zhu H, Cundiff P, Pietrusiak P, Edwards S, Phillips JW, Sachs A, Schadt EE (2004) Genetic inheritance of gene expression in human cell lines. *Am J Hum Genet* 75:1094–1105
55. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, Cheung VG (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature* 430:743–7
56. Ooi SL, Pan X, Peyser BD, Ye P, Meluh PB, Yuan DS, Irizarry RA, Bader JS, Spencer FA, Boeke JD (2006) Global synthetic-lethality analysis and yeast functional profiling. *Trends Genet* 22:56–63
57. Pan X, Ye P, Yuan DS, Wang X, Bader JS, Boeke JD (2006) A DNA integrity network in the yeast *Saccharomyces cerevisiae*. *Cell* 124:1069–81
58. Pan X, Yuan DS, Xiang D, Wang X, Sookhai-Mahadeo S, Bader JS, Hieter P, Spencer F, Boeke JD (2004) A robust toolkit for functional profiling of the yeast genome. *Mol Cell* 16:487–96
59. Perlstein EO, Ruderfer DM, Roberts DC, Schreiber SL, Kruglyak L (2007) Genetic basis of individual differences in the response to small-molecule drugs in yeast. *Nat Genet* 39:496–502
60. Ramsey SA, Smith JJ, Orrell D, Marelli M, Petersen TW, de Atauri P, Bolouri H, Aitchison JD (2006) Dual feedback loops in the GAL regulon suppress cellular heterogeneity in yeast. *Nat Genet* 38:1082–1087
61. Reguly T, Breitkreutz A, Boucher L, Breitkreutz BJ, Hon GC, Myers CL, Parsons A, Friesen H, Oughtred R, Tong A, Stark C, Ho Y, Botstein D, Andrews B, Boone C, Troyanskaya OG, Ideker T, Dolinski K, Batada NN, Tyers M (2006) Comprehensive curation and analysis of global interaction networks in *Saccharomyces cerevisiae*. *J Biol* 5:11
62. Rockman MV, Kruglyak L (2006) Genetics of global gene expression. *Nat Rev Genet* 7:862–72
63. Schacherer J, Ruderfer DM, Gresham D, Dolinski K, Botstein D, Kruglyak L (2007) Genome-wide analysis of nucleotide-level variation in commonly used *Saccharomyces cerevisiae* strains. *PLoS ONE* 2:e322
64. Schuldiner M, Collins SR, Thompson NJ, Denic V, Bhamidipati A, Punna T, Ihmels J, Andrews B, Boone C, Greenblatt JF, Weissman JS, Krogan NJ (2005) Exploration of the function and organization of the yeast early secretory pathway through an epistatic miniarray profile. *Cell* 123:507–19
65. Segal E, Shapira M, Regev A, Pe'er D, Botstein D, Koller D, Friedman N (2003) Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet* 34:166–76
66. Segre D, Deluna A, Church GM, Kishony R (2005) Modular epistasis in yeast metabolism. *Nat Genet* 37:77–83
67. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 13:2498–504
68. Shmulevich I, Kauffman SA, Aldana M (2005) Eukaryotic cells are dynamically ordered or critical but not chaotic. *Proc Natl Acad Sci USA* 102:13439–44
69. Sinha H, Nicholson BP, Steinmetz LM, McCusker JH (2006) Complex genetic interactions in a quantitative trait locus. *PLoS Genet* 2:e13
70. Spiro PA, Parkinson JS, Othmer HG (1997) A model of excitation and adaptation in bacterial chemotaxis. *Proc Natl Acad Sci USA* 94:7263–8
71. St Onge RP, Mani R, Oh J, Proctor M, Fung E, Davis RW, Nislow C, Roth FP, Giaever G (2007) Systematic pathway analysis using high-resolution fitness profiling of combinatorial gene deletions. *Nat Genet* 39:199–206
72. Steinmetz LM, Sinha H, Richards DR, Spiegelman JJ, Oefner PJ, McCusker JH, Davis RW (2002) Dissecting the architecture of a quantitative trait locus in yeast. *Nature* 416:326–30
73. Stelzl U, Worm U, Lalowski M, Haenig C, Brembeck FH, Goehler H, Stroedicke M, Zenkner M, Schoenherr A, Koeppen S, Timm J, Mintzlaff S, Abraham C, Bock N, Kietzmann S, Goedde A, Toksoz E, Droegge A, Krobitsch S, Korn B, Birchmeier W, Lehrach H, Wanker EE (2005) A human protein-protein interaction network: a resource for annotating the proteome. *Cell* 122:957–68

74. Storey JD, Akey JM, Kruglyak L (2005) Multiple locus linkage analysis of genomewide expression in yeast. *PLoS Biol* 3:e267
75. Sturtevant AH (1956) A Highly Specific Complementary Lethal System in *Drosophila Melanogaster*. *Genetics* 41:118–23
76. Taylor RJ, Siegel AF, Galitski T (2007) Network motif analysis of a multi-mode genetic-interaction network. *Genome Biol* 8:R160
77. Thakar J, Pilione M, Kirimanjeswara G, Harvill ET, Albert R (2007) Modeling systems-level regulation of host immune responses. *PLoS Comput Biol* 3:e109
78. Tischler J, Lehner B, Chen N, Fraser AG (2006) Combinatorial RNA interference in *Caenorhabditis elegans* reveals that redundancy between gene duplicates can be maintained for more than 80 million years of evolution. *Genome Biol* 7:R69
79. Tong AH, Evangelista M, Parsons AB, Xu H, Bader GD, Page N, Robinson M, Raghibizadeh S, Hogue CW, Bussey H, Andrews B, Tyers M, Boone C (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* 294:2364–2368
80. Tong AH, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu AM, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C (2004) Global mapping of the yeast genetic interaction network. *Science* 303:808–13
81. Tucker CL, Fields S (2003) Lethal combinations. *Nat Genet* 35:204–5
82. Van Driessche N, Demsar J, Booth EO, Hill P, Juvan P, Zupan B, Kuspa A, Shaulsky G (2005) Epistasis analysis with global transcriptional phenotypes. *Nat Genet* 37:471–7
83. Vidal M (2005) Interactome modeling. *FEBS Lett* 579:1834–8
84. Waddington CH (1942) Canalization of Development and the Inheritance of Acquired Characters. *Nature* 150:563
85. Wayne ML, McIntyre LM (2002) Combining mapping and arraying: An approach to candidate gene identification. *Proc Natl Acad Sci USA* 99:14903–6
86. Winston F, Dollard C, Ricupero-Hovasse SL (1995) Construction of a set of convenient *Saccharomyces cerevisiae* strains that are isogenic to S288C. *Yeast* 11:53–5
87. Winzler EA, Shoemaker DD, Astromoff A, Liang H, Anderson K, Andre B, Bangham R, Benito R, Boeke JD, Bussey H, Chu AM, Connelly C, Davis K, Dietrich F, Dow SW, El Bakkoury M, Foury F, Friend SH, Gentalen E, Giaever G, Hegemann JH, Jones T, Laub M, Liao H, Liebundguth N, Lockhart DJ, Lucau-Danila A, Lussier M, M'Rabet N, Menard P, Mittmann M, Pai C, Rebischung C, Revuelta JL, Riles L, Roberts CJ, Ross-MacDonald P, Scherens B, Snyder M, Sookhai-Mahadeo S, Storms RK, Veronneau S, Voet M, Volckaert G, Ward TR, Wysocki R, Yen GS, Yu K, Zimmermann K, Philippsen P, Johnston M, Davis RW (1999) Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* 285:901–6
88. Wong SL, Zhang LV, Tong AH, Li Z, Goldberg DS, King OD, Lesage G, Vidal M, Andrews B, Bussey H, Boone C, Roth FP (2004) Combining biological networks to predict genetic interactions. *Proc Natl Acad Sci USA* 101:15682–7
89. Workman CT, Mak HC, McCuine S, Tagne JB, Agarwal M, Ozier O, Begley TJ, Samson LD, Ideker T (2006) A systems approach to mapping DNA damage response pathways. *Science* 312:1054–1059
90. Ye P, Peyser BD, Pan X, Boeke JD, Spencer FA, Bader JS (2005) Gene function prediction from congruent synthetic lethal interactions in yeast. *Mol Syst Biol* 1:0026
91. Yeang CH, Mak HC, McCuine S, Workman C, Jaakkola T, Ideker T (2005) Validation and refinement of gene-regulatory pathways on a network of physical interactions. *Genome Biol* 6:R62
92. Yeger-Lotem E, Sattath S, Kashtan N, Itzkovitz S, Milo R, Pinter RY, Alon U, Margalit H (2004) Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc Natl Acad Sci USA* 101:5934–9
93. Yeh P, Tschumi AI, Kishony R (2006) Functional classification of drugs by properties of their pairwise interactions. *Nat Genet* 38:489–94
94. Zhang LV, King OD, Wong SL, Goldberg DS, Tong AH, Lesage G, Andrews B, Bussey H, Boone C, Roth FP (2005) Motifs, themes and thematic maps of an integrated *Saccharomyces cerevisiae* interaction network. *J Biol* 4:6
95. Zhong W, Sternberg PW (2006) Genome-wide prediction of *C. elegans* genetic interactions. *Science* 311:1481–4